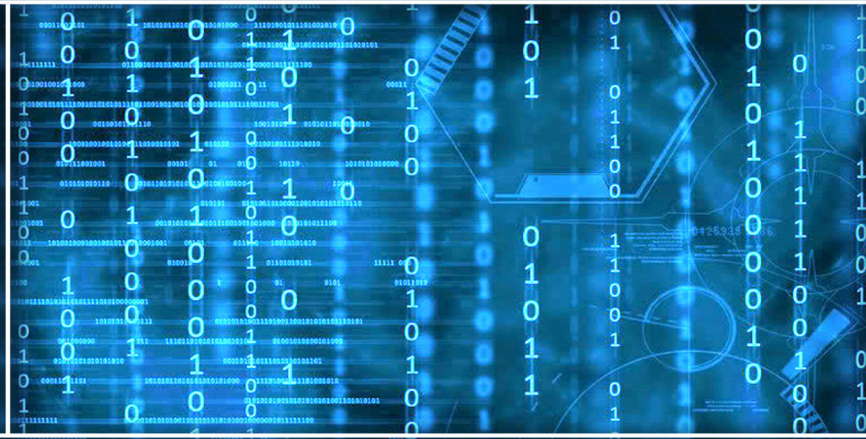


Volume 13 Issue 9

September 2022



ISSN 2156-5570(Online)

ISSN 2158-107X(Print)



# Editorial Preface

## *From the Desk of Managing Editor...*

It may be difficult to imagine that almost half a century ago we used computers far less sophisticated than current home desktop computers to put a man on the moon. In that 50 year span, the field of computer science has exploded.

Computer science has opened new avenues for thought and experimentation. What began as a way to simplify the calculation process has given birth to technology once only imagined by the human mind. The ability to communicate and share ideas even though collaborators are half a world away and exploration of not just the stars above but the internal workings of the human genome are some of the ways that this field has moved at an exponential pace.

At the International Journal of Advanced Computer Science and Applications it is our mission to provide an outlet for quality research. We want to promote universal access and opportunities for the international scientific community to share and disseminate scientific and technical information.

We believe in spreading knowledge of computer science and its applications to all classes of audiences. That is why we deliver up-to-date, authoritative coverage and offer open access of all our articles. Our archives have served as a place to provoke philosophical, theoretical, and empirical ideas from some of the finest minds in the field.

We utilize the talents and experience of editor and reviewers working at Universities and Institutions from around the world. We would like to express our gratitude to all authors, whose research results have been published in our journal, as well as our referees for their in-depth evaluations. Our high standards are maintained through a double blind review process.

We hope that this edition of IJACSA inspires and entices you to submit your own contributions in upcoming issues. Thank you for sharing wisdom.

**Thank you for Sharing Wisdom!**

**Kohei Arai**  
**Editor-in-Chief**  
**IJACSA**  
**Volume 13 Issue 9 September 2022**  
**ISSN 2156-5570 (Online)**  
**ISSN 2158-107X (Print)**



# Editorial Board

## Editor-in-Chief

### **Dr. Kohei Arai - Saga University**

*Domains of Research: Technology Trends, Computer Vision, Decision Making, Information Retrieval, Networking, Simulation*

---

## Associate Editors

### **Alaa Sheta**

#### **Southern Connecticut State University**

*Domain of Research: Artificial Neural Networks, Computer Vision, Image Processing, Neural Networks, Neuro-Fuzzy Systems*

### **Domenico Ciuonzo**

#### **University of Naples, Federico II, Italy**

*Domain of Research: Artificial Intelligence, Communication, Security, Big Data, Cloud Computing, Computer Networks, Internet of Things*

### **Dorota Kaminska**

#### **Lodz University of Technology**

*Domain of Research: Artificial Intelligence, Virtual Reality*

### **Elena Scutelnicu**

#### **"Dunarea de Jos" University of Galati**

*Domain of Research: e-Learning, e-Learning Tools, Simulation*

### **In Soo Lee**

#### **Kyungpook National University**

*Domain of Research: Intelligent Systems, Artificial Neural Networks, Computational Intelligence, Neural Networks, Perception and Learning*

### **Krassen Stefanov**

#### **Professor at Sofia University St. Kliment Ohridski**

*Domain of Research: e-Learning, Agents and Multi-agent Systems, Artificial Intelligence, e-Learning Tools, Educational Systems Design*

### **Renato De Leone**

#### **Università di Camerino**

*Domain of Research: Mathematical Programming, Large-Scale Parallel Optimization, Transportation problems, Classification problems, Linear and Integer Programming*

### **Xiao-Zhi Gao**

#### **University of Eastern Finland**

*Domain of Research: Artificial Intelligence, Genetic Algorithms*

# CONTENTS

**Paper 1: ModER: Graph-based Unsupervised Entity Resolution using Composite Modularity Optimization and Locality Sensitive Hashing**

*Authors: Islam Akef Ebeid, John R. Talburt, Nicholas Kofi Akortia Hagan, Md Abdus Salam Siddique*

**PAGE 1 – 18**

**Paper 2: Remote International Collaboration in Scientific Research Teams for Technology Development**

*Authors: Sarah Janböcke, Toshimi Ogawa, Koki Kobayashi, Ryan Browne, Yasuki Taki, Rainer Wieching, Johanna Langendorf*

**PAGE 19 – 29**

**Paper 3: Fuzzy Image Enhancement Method based on a New Intensifier Operator**

*Authors: Libao Yang, Suzelawati Zenian, Rozaimi Zakaria*

**PAGE 30 – 34**

**Paper 4: Cooperative Multi-Robot Hierarchical Reinforcement Learning**

*Authors: Gembong Edhi Setyawan, Pitoyo Hartono, Hideyuki Sawada*

**PAGE 35 – 44**

**Paper 5: Differential Privacy Technology of Big Data Information Security based on ACA-DMLP**

*Authors: Yubiao Han, Lei Wang, Dianhong He*

**PAGE 45 – 52**

**Paper 6: A Reusable Product Line Asset in Smart Mobile Application: A Systematic Literature Review**

*Authors: Nan Pepin, Abdul S. Shibghatullah, Kasthuri Subaramaniam, Rabatul Aduni Sulaiman, Zuraida A. Abas, Samer Sarsam*

**PAGE 53 – 60**

**Paper 7: A Study on the Effect of Digital Fabrication in Social Studies Education**

*Authors: Kazunari Hirakoso, Hidetake Hamada*

**PAGE 61 – 66**

**Paper 8: A Blockchain-based Model for Securing IoT Transactions in a Healthcare Environment**

*Authors: Mohamed Abdel Kader Mohamed Elgendy, Mohamed Aborizka, Ali Mohamed Nabil Allam*

**PAGE 67 – 75**

**Paper 9: Study on Early Warning on the Financial Risk of Project Venture Capital through a Neural Network Model**

*Authors: Xianjuan Li*

**PAGE 76 – 81**

**Paper 10: Improving Privacy Preservation Approach for Healthcare Data using Frequency Distribution of Delicate Information**

*Authors: Ganesh Dagadu Puri, D. Haritha*

**PAGE 82 – 90**

**Paper 11: Attention-based Long Short Term Memory Model for DNA Damage Prediction in Mammalian Cells**

*Authors: Mohammad A. Alsharaiah, Laith H. Baniata, Omar Adwan, Ahmad Adel Abu-Shareha, Mosleh Abu Alhaj, Qasem Kharma, Abdelrahman Hussein, Orieb Abualghanam, Nabeel Alassaf, Mohammad Baniata*

**PAGE 91 – 99**

**Paper 12: Estimation of Recovery Percentage in Gravimetric Concentration Processes using an Artificial Neural Network Model**

*Authors: Manuel Alejandro Ospina-Alarcón, Ismael E. Rivera-M, Gabriel Elías Chanchí-Golondrino*

**PAGE 100 – 110**

**Paper 13: Risk Prediction Applied to Global Software Development using Machine Learning Methods**

*Authors: Hossam Hassan, Manal A. Abdel-Fattah, Amr Ghoneim*

**PAGE 111 – 120**

**Paper 14: HelaNER 2.0: A Novel Deep Neural Model for Named Entity Boundary Detection**

*Authors: Y. H. P. P Priyadarshana, L Ranathunga*

**PAGE 121 – 130**

**Paper 15: Face Recognition under Illumination based on Optimized Neural Network**

*Authors: Napa Lakshmi, Megha P Arakeri*

**PAGE 131 – 137**

**Paper 16: Transfer Learning for Medicinal Plant Leaves Recognition: A Comparison with and without a Fine-Tuning Strategy**

*Authors: Vina Ayumi, Ermatita Ermatita, Abdiansah Abdiansah, Handrie Noprisson, Yuwan Jumaryadi, Mariana Purba, Marissa Utami, Erwin Dwika Putra*

**PAGE 138 – 144**

**Paper 17: Effect of Random Splitting and Cross Validation for Indonesian Opinion Mining using Machine Learning Approach**

*Authors: Mariana Purba, Ermatita Ermatita, Abdiansah Abdiansah, Handrie Noprisson, Vina Ayumi, Hadiguna Setiawan, Umniy Salamah, Yadi Yadi*

**PAGE 145 – 151**

**Paper 18: Classification of Diabetes Types using Machine Learning**

*Authors: Oyeranmi Adigun, Folasade Okikiola, Nureni Yekini, Ronke Babatunde*

**PAGE 152 – 161**

**Paper 19: Criteria and Guideline for Dyslexic Intervention Games**

*Authors: Noraziah ChePa, Nur Azzah Abu Bakar, Laura Lim Sie-Yi*

**PAGE 162 – 172**

**Paper 20: The Effectiveness of Gamification for Students' Engagement in Technical and Vocational Education and Training**

*Authors: Laily Abu Samah, Amirah Ismail, Mohammad Kamrul Hasan*

**PAGE 173 – 180**

**Paper 21: Campus Quality of Services Analysis of Mobile Wireless Communications Network Signal among Providers in Malaysia**

*Authors: Murizah Kassim, Zulfadhli Hisam, Mohd Nazri Ismail*

**PAGE 181 – 187**

**Paper 22: Exploring Alumni Data using Data Visualization Techniques**

*Authors: Nurhanani Izzati Ismail, Nur Atiqah Sia Abdullah, Nasiroh Omar*

**PAGE 188 – 195**



Paper 23: The Performance Evaluation of Transfer Learning VGG16 Algorithm on Various Chest X-ray Imaging Datasets for COVID-19 Classification

*Authors: Andi Sunyoto, Yoga Pristyanto, Arief Setyanto, Fawaz Alarfaj, Naif Almusallam, Mohammed Alreshoodi*

PAGE 196 – 203

Paper 24: A Comprehensive Review and Application of Interpretable Deep Learning Model for ADR Prediction

*Authors: Shiksha Alok Dubey, Anala A. Pandit*

PAGE 204 – 213

Paper 25: Secure Cloud Connected Indoor Hydroponic System via Multi-factor Authentication

*Authors: Mohamad Khairul Hafizi Rahimi, Mohamad Hanif Md Saad, Aini Hussain, Nurul Maisarah Hamdan*

PAGE 214 – 222

Paper 26: Effective Multitier Network Model for MRI Brain Disease Prediction using Learning Approaches

*Authors: N. Ravinder, Moulana Mohammed*

PAGE 223 – 230

Paper 27: Application based on Hybrid CNN-SVM and PCA-SVM Approaches for Classification of Cocoa Beans

*Authors: AYIKPA Kacoutchy Jean, MAMADOU Diarra, BALLO Abou Bakary, GOUTON Pierre, ADOU Kablan Jérôme*

PAGE 231 – 238

Paper 28: SQrum: An Improved Method of Scrum

*Authors: Najih Soukaina, Merzouk Soukaina, Marzak Abdelaziz*

PAGE 239 – 249

Paper 29: Use of Interactive Multimedia e-Learning in TVET Education

*Authors: Siti Fadzilah Mat Noor, Hazura Mohamed, Nur Atiqah Zaini, Dayana Daiman*

PAGE 250 – 256

Paper 30: CBT4Depression: A Cognitive Behaviour Therapy (CBT) Therapeutic Game to Reduce Depression Level among Adolescents

*Authors: Norhana Yusof, Nazrul Azha Mohamed Shaari, Eizwan Hamdie Yusoff*

PAGE 257 – 264

Paper 31: Creating Video Visual Storyboard with Static Video Summarization using Fractional Energy of Orthogonal Transforms

*Authors: Ashvini Tonge, Sudeep D. Thepade*

PAGE 265 – 273

Paper 32: Denoising of Impulse Noise using Partition-Supported Median, Interpolation and DWT in Dental X-Ray Images

*Authors: Mohamed Shajahan, Siti Armiza Mohd Aris, Sahnus Usman, Norliza Mohd Noor*

PAGE 274 – 280

Paper 33: An End-to-End Big Data Deduplication Framework based on Online Continuous Learning

*Authors: Widad Elouataoui, Imane El Alaoui, Saida El Mendili, Youssef Gahi*

PAGE 281 – 291

Paper 34: Student's Performance Prediction based on Personality Traits and Intelligence Quotient using Machine Learning

*Authors: Samar El-Keiey, Dina ElMenshawy, Ehab Hassanein*

PAGE 292 – 299

Paper 35: Real Time Fire Detection using Color Probability Segmentation and DenseNet Model for Classifier

*Authors: Faisal Dharma Adhinata, Nur Ghaniaviyanto Ramadhan*

**PAGE 300 – 305**

Paper 36: Tissue and Tumor Epithelium Classification using Fine-tuned Deep CNN Models

*Authors: Anju T E, S. Vimala*

**PAGE 306 – 314**

Paper 37: Predicting University Student Retention using Artificial Intelligence

*Authors: Samer M. Arqawi, Eman Akef Zitawi, Anees Husni Rabaya, Basem S. Abunasser, Samy S. Abu-Naser*

**PAGE 315 – 324**

Paper 38: Using the Agglomerative Hierarchical Clustering Method to Examine Human Factors in Indonesian Aviation Accidents

*Authors: Rossi Passarella, Gulfi Oktariani, Dedy Kurniawan, Purwita Sari*

**PAGE 325 – 331**

Paper 39: A Framework for Crime Detection and Diminution in Digital Forensics (CD3F)

*Authors: Arpita Singh, Sanjay K. Singh, Hari Kiran Vege, Nilu Singh*

**PAGE 332 – 345**

Paper 40: Deep Learning and Classification Algorithms for COVID-19 Detection

*Authors: Mohammed Sidheeque, P. Sumathy, Abdul Gafur. M*

**PAGE 346 – 350**

Paper 41: Gamification on OTT Platforms: A Behavioural Study for User Engagement

*Authors: Komal Suryavanshi, Prasun Gahlot, Surya Bahadur Thapa, Aradhana Gandhi, Ramakrishnan Raman*

**PAGE 351 – 363**

Paper 42: Optimally Allocating Ambulances in Delhi using Mutation based Shuffled Frog Leaping Algorithm

*Authors: Zaheeruddin, Hina Gupta*

**PAGE 364 – 374**

Paper 43: Adopting a Digital Transformation in Moroccan Research Structure using a Knowledge Management System: Case of a Research Laboratory

*Authors: Fatima-Ezzahra AIT-BENNACER, Abdessadek AAROUD, Khalid AKODADI, Bouchaib CHERRADI*

**PAGE 375 – 384**

Paper 44: Analyzing the Relationship between the Personality Traits and Drug Consumption (Month-based user Definition) using Rough Sets Theory

*Authors: Manasik M. Nour, H. A. Mohamed, Sumayyah I. Alshber*

**PAGE 385 – 392**

Paper 45: Wavelet Multi Resolution Analysis based Data Hiding with Scanned Secrete Images

*Authors: Kohei Arai*

**PAGE 393 – 400**

Paper 46: Multiple Eye Disease Detection using Hybrid Adaptive Mutation Swarm Optimization and RNN

*Authors: P. Glaret Subin, P. Muthu Kannan*

**PAGE 401 – 410**

**Paper 47: Insect Pest Image Detection and Classification using Deep Learning**

*Authors: Niranjan C Kundur, P B Mallikarjuna*

**PAGE 411 – 421**

**Paper 48: Analysis of Noise Removal Techniques on Retinal Optical Coherence Tomography Images**

*Authors: T. M. Sheeba, S. Albert Antony Raj*

**PAGE 422 – 427**

**Paper 49: Analyzing the State of Mind of Self-quarantined People during COVID-19 Pandemic Lockdown Period: A Multiple Correspondence Analysis Approach**

*Authors: Gauri Vaidya, Vidya Kumbhar, Sachin Naik, Vijayatai Hukare*

**PAGE 428 – 439**

**Paper 50: SIBI (Sign System Indonesian Language) Text-to-3D Animation Translation Mobile Application**

*Authors: Erdefi Rakun, Sultan Muzahidin, IGM Surya A. Darmana, Wikan Setiaji*

**PAGE 440 – 450**

**Paper 51: A Review of Foreground Segmentation based on Convolutional Neural Networks**

*Authors: Pavan Kumar Tadiparthi, Sagarika Bugatha, Pradeep Kumar Bheemavarapu*

**PAGE 451 – 454**

**Paper 52: Multi-instance Finger Knuckle Print Recognition based on Fusion of Local Features**

*Authors: Amine AMRAOUI, Mounir AIT KERROUM, Youssef FAKHRI*

**PAGE 455 – 463**

**Paper 53: Detection of Credit Card Fraud using a Hybrid Ensemble Model**

*Authors: Sayali Saraf, Anupama Phakatkar*

**PAGE 464 – 474**

**Paper 54: Covid-19 and Pneumonia Infection Detection from Chest X-Ray Images using U-Net, EfficientNetB1, XGBoost and Recursive Feature Elimination**

*Authors: Munindra Lunagaria, Vijay Katkar, Krunal Vaghela*

**PAGE 475 – 483**

**Paper 55: Analysis of Privacy and Security Challenges in e-Health Clouds**

*Authors: Reem Alanazi*

**PAGE 484 – 489**

**Paper 56: Identification of Retinal Disease using Anchor-Free Modified Faster Region**

*Authors: Arulselvam. T, S. J. Sathish Aaron Joseph*

**PAGE 490 – 499**

**Paper 57: OpenCV Implementation of Grid-based Vertical Safe Landing for UAV using YOLOv5**

*Authors: Hrusna Chakri Shadakshri V, Veena M. B, Keshihaa Rudra Gana Dev V*

**PAGE 500 – 506**

**Paper 58: Gaussian Projection Deep Extreme Clustering and Chebyshev Reflective Correlation based Outlier Detection**

*Authors: S. Rajalakshmi, P. Madhubala*

**PAGE 507 – 515**



Paper 59: Efficient Decentralized Sharing Economy Model based on Blockchain Technology: A Case Study of Najm for Insurance Services Company

*Authors: Atheer Alkhamash, Kawther Saeedi, Fatmah Baothman, Rania Anwar Aboalela, Amal Babour*

PAGE 516 – 522

Paper 60: Virtual Communities of Practice to Promote Digital Agriculturists' Learning Competencies and Learning Engagement: Conceptual Framework

*Authors: Maneerat Manyuen, Surapon Boonlue, Jariya Neanchaleay, Vitsanu Nittayathamkul*

PAGE 523 – 528

Paper 61: Deep Q-learning Approach based on CNN and XGBoost for Traffic Signal Control

*Authors: Nada Faqir, Chakir Loqman, Jaouad Boumhidi*

PAGE 529 – 536

Paper 62: Automatic Text Summarization using Document Clustering Named Entity Recognition

*Authors: Senthamizh Selvan. R, K. Arutchelvan*

PAGE 537 – 543

Paper 63: Convolutional Neural Networks with Transfer Learning for Pneumonia Detection

*Authors: Orlando Iparraguirre-Villanueva, Victor Guevara-Ponce, Ofelia Roque Paredes, Fernando Sierra-Liñan, Joselyn Zapata-Paulini, Michael Cabanillas-Carbonell*

PAGE 544 – 551

Paper 64: A Monadic Co-simulation Model for Cyber-physical Production Systems

*Authors: Daniel-Cristian Crăciunean*

PAGE 552 – 557

Paper 65: A Machine Learning Model for Predicting Heart Disease using Ensemble Methods

*Authors: Jasjit Singh Samagh, Dilbag Singh*

PAGE 558 – 565

Paper 66: Novel Approach in Classification and Prediction of COVID-19 from Radiograph Images using CNN

*Authors: Chalapathiraju Kanumuri, CH. Renu Madhavi, Torthi Ravichandra*

PAGE 566 – 570

Paper 67: Machine Learning based Electromigration-aware Scheduler for Multi-core Processors

*Authors: Jagadeesh Kumar P, Mini M G*

PAGE 571 – 580

Paper 68: Sentiment Analysis on Acceptance of New Normal in COVID-19 Pandemic using Naïve Bayes Algorithm

*Authors: Siti Hajar Aishah Samsudin, Norlina Mohd Sabri, Norulhidayah Isa, Ummu Fatimah Mohd Bahrin*

PAGE 581 – 588

Paper 69: Partial Differential Equation (PDE) based Hybrid Diffusion Filters for Enhancing Noise Performance of Point of Care Ultrasound (POCUS) Images

*Authors: Deepa V S, Jagathyraj V P, Gopikakumari R*

PAGE 589 – 596

Paper 70: Smart Greenhouse Monitoring and Controlling based on NodeMCU

*Authors: Yajie Liu*

PAGE 597 – 600

Paper 71: Design of Accounting Information System in Data Processing: Case Study in Indonesia Company  
Authors: Meiryani, Dezie Leonarda Warganegara, Agustinus Winoto, Gabrielle Beatrice Hidayat, Erna Bernadetta Sitanggang, Ka Tiong, Jessica Paulina Sidauruk, Mochammad Fahlevi, Gredion Prajena  
PAGE 601 – 606

Paper 72: MOOC Dropout Prediction using FIAR-ANN Model based on Learner Behavioral Features  
Authors: S. Nithya, S.Umarani  
PAGE 607 – 617

Paper 73: Sentiment Analysis of Online Movie Reviews using Machine Learning  
Authors: Isaiah Steinke, Justin Wier, Lindsay Simon, Raed Seetan  
PAGE 618 – 624

Paper 74: Detection and Extraction of Faces and Text Lower Third Techniques for an Audiovisual Archive System using Machine Learning  
Authors: Khalid El Fayq, Said Tkatek, Lahcen Idouglid, Jaafar Abouchabaka  
PAGE 625 – 632

Paper 75: Data Recovery Comparative Analysis using Open-based Forensic Tools Source on Linux  
Authors: Muhammad Fahmi Abdillah, Yudi Prayudi  
PAGE 633 – 639

Paper 76: Advanced Persistent Threat Attack Detection using Clustering Algorithms  
Authors: Ahmed Alsanad, Sara Altuwaijri  
PAGE 640 – 649

Paper 77: Energy Efficient Node Deployment Technique for Heterogeneous Wireless Sensor Network based Object Detection  
Authors: Jayashree Dev, Jibitesh Mishra  
PAGE 650 – 659

Paper 78: Fine-grained Access Control in Distributed Cloud Environment using Trust Valuation Model  
Authors: Aparna Manikonda, Nalini N  
PAGE 660 – 666

Paper 79: BERT-Based Hybrid RNN Model for Multi-class Text Classification to Study the Effect of Pre-trained Word Embeddings  
Authors: Shreyashree S, Pramod Sunagar, S Rajarajeswari, Anita Kanavalli  
PAGE 667 – 674

Paper 80: A Hybrid Approach of Wavelet Transform, Convolutional Neural Networks and Gated Recurrent Units for Stock Liquidity Forecasting  
Authors: Mohamed Ben Houad, Mohammed Mestari, Khalid Bentaleb, Adnane El Mansouri, Salma El Aidouni  
PAGE 675 – 682

Paper 81: Visual Navigation System for Autonomous Drone using Fiducial Marker Detection  
Authors: Mohammad Soleimani Amiri, Rizauddin Ramli  
PAGE 683 – 690

**Paper 82: Design of a Mobile Application for the Logistics Process of a Fire Company**

*Authors: Luis Enrique Parra Aquije, Luis Gustavo Vasquez Carranza, Gustavo Bernnet Alfaro Pena, Michael Cabanillas-Carbonell, Laberiano Andrade-Arenas*

**PAGE 691 – 699**

**Paper 83: Intelligent System for Personalised Interventions and Early Drop-out Prediction in MOOCs**

*Authors: ALJ Zakaria, BOUAYAD Anas, Cherkaoui Malki Mohammed Oucamah*

**PAGE 700 – 710**

**Paper 84: An Intelligent Decision Support Ensemble Voting Model for Coronary Artery Disease Prediction in Smart Healthcare Monitoring Environments**

*Authors: Anas Maach, Jamila Elalami, Noureddine Elalami, El Houssine El Mazoudi*

**PAGE 711 – 724**

**Paper 85: Estimation of Varying Reaction Times with RNN and Application to Human-like Autonomous Car-following Modeling**

*Authors: Lijing Ma, Shiru Qu, Junxi Zhang, Xiangzhou Zhang*

**PAGE 725 – 730**

**Paper 86: Mobile Food Journalling Application with Convolutional Neural Network and Transfer Learning: A Case for Diabetes Management in Malaysia**

*Authors: Jason Thomas Chew, Yakub Sebastian, Valliapan Raman, Patrick Hang Hui Then*

**PAGE 731 – 737**

**Paper 87: Rethinking Classification of Oriented Object Detection in Aerial Images**

*Authors: Phuc Nguyen, Thang Truong, Nguyen D. Vo, Khang Nguyen*

**PAGE 738 – 747**

**Paper 88: TextBrew: Automated Model Selection and Hyperparameter Optimization for Text Classification**

*Authors: Rushil Desai, Aditya Shah, Shourya Kothari, Aishwarya Surve, Narendra Shekokar*

**PAGE 748 – 754**

**Paper 89: On-Device Major Indian Language Identification Classifiers to Run on Low Resource Devices**

*Authors: Yashwanth Y S*

**PAGE 755 – 760**

**Paper 90: Evaluating Hybrid Framework of VASNET and IoT in Disaster Management System**

*Authors: Sia Chiu Shoon, Mohammad Nazim Jambli, Sinarwati Mohamad Suhaili, Nur Haryani Zakaria*

**PAGE 761 – 766**

**Paper 91: A Novel Machine Learning-based Framework for Detecting Religious Arabic Hatred Speech in Social Networks**

*Authors: Mahmoud Masadeh, Hanumanthappa Jayappa Davanager, Abdullah Y. Muaad*

**PAGE 767 – 776**

**Paper 92: Modeling Multioutput Response Uses Ridge Regression and MLP Neural Network with Tuning Hyperparameter through Cross Validation**

*Authors: Waego Hadi Nugroho, Samingun Handoyo, Hsing-Chuan Hsieh, Yusnita Julyarni Akri, Zuraidah, Donna DwinitaAdelia*

**PAGE 777 – 787**



**Paper 93: Decentralized Access Control using Blockchain Technology for Application in Smart Farming**

*Authors: Normaizeerah Mohd Noor, Noor Afiza Mat Razali, Nur Atiqah Malizan, Khairul Khalil Ishak, Muslihah Wook, Nor Asiakin Hasbullah*

**PAGE 788 – 802**

**Paper 94: Research on Intelligent Control System of Air Conditioning based on Internet of Things Intelligent Control System of Air Conditioning**

*Authors: Binfang Zhang*

**PAGE 803 – 814**

**Paper 95: A Short Review on the Role of Various Deep Learning Techniques for Segmenting and Classifying Brain Tumours from MRI Images**

*Authors: Kumari Kavitha. D, E. Kiran Kumar*

**PAGE 815 – 824**

**Paper 96: Fish Species Classification using Optimized Deep Learning Model**

*Authors: J. M. Jini Mol, S. Albin Jose*

**PAGE 825 – 837**

**Paper 97: Environmental Noise Pollution Forecasting using Fuzzy-autoregressive Integrated Moving Average Modelling**

*Authors: Muhammad Shukri Che Lah, Nureize Arbaiy, Syahir Ajwad Sapuan, Pei-Chun Lin*

**PAGE 838 – 843**

**Paper 98: Extractive Multi-document Text Summarization Leveraging Hybrid Semantic Similarity Measures**

*Authors: Rajesh Bandaru, Y. Radhika*

**PAGE 844 – 852**

**Paper 99: An Efficient Hybrid LSTM-CNN and CNN-LSTM with GloVe for Text Multi-class Sentiment Classification in Gender Violence**

*Authors: Abdul Azim Ismail, Marina Yusoff*

**PAGE 853 – 863**

**Paper 100: Performance Analysis of Deep Learning YOLO models for South Asian Regional Vehicle Recognition**

*Authors: Minar Mahmud Rafi, Siddharth Chakma, Asif Mahmud, Raj Xavier Rozario, Rukon Uddin Munna, Md. Abrar Abedin Wohra, Rakibul Haque Joy, Khan Raqib Mahmud, Bijan Paul*

**PAGE 864 – 873**

**Paper 101: Multi-method Approach for User Experience of Selfie-taking Mobile Applications**

*Authors: Shahad Aldahri, Reem Alnanih*

**PAGE 874 – 880**

**Paper 102: Predicting Academic Performance using a Multiclassification Model: Case Study**

*Authors: Alfredo Daza Vergaray, Carlos Guerra, Noemi Cervera, Erwin Burgos*

**PAGE 881 – 889**

**Paper 103: COVID-19 Disease Detection based on X-Ray Image Classification using CNN with GEV Activation Function**

*Authors: Karim Ali Mohamed, Emad Elsamahy, Ahmed Salem*

**PAGE 890 – 898**

Paper 104: Deep Learning based Cervical Cancer Classification and Segmentation from Pap Smears Images using an EfficientNet

Authors: Krishna Prasad Battula, B. Sai Chandana

PAGE 899 – 908

Paper 105: Cloud based Forecast of Municipal Solid Waste Growth using AutoRegressive Integrated Moving Average Model: A Case Study for Bengaluru

Authors: Rashmi G, S Sathish Kumar K

PAGE 909 – 913

Paper 106: Building an Intelligent Tutoring System for Learning Polysemous Words in Moore

Authors: Pengwende ZONGO, Tounwendyam Frederic OUEDRAOGO

PAGE 914 – 920

Paper 107: Improving the Diabetes Diagnosis Prediction Rate Using Data Preprocessing, Data Augmentation and Recursive Feature Elimination Method

Authors: E. Sabitha, M. Durgadevi

PAGE 921 – 930

Paper 108: A Comparative Study of Unsupervised Anomaly Detection Algorithms used in a Small and Medium-Sized Enterprise

Authors: Irina Petrariu, Adrian Moscaliuc, Cristina Elena Turcu, Ovidiu Gherman

PAGE 931 – 940

Paper 109: Automated Brain Disease Classification using Transfer Learning based Deep Learning Models

Authors: Farhana Alam, Farhana Chowdhury Tisha, Sara Anisa Rahman, Samia Sultana, Md. Ahied Mahi Chowdhury, Ahmed Wasif Reza, Mohammad Shamsul Arefin

PAGE 941 – 949

Paper 110: Toward A Holistic, Efficient, Stacking Ensemble Intrusion Detection System using a Real Cloud-based Dataset

Authors: Ahmed M. Mahfouz, Abdullah Abuhussein, Faisal S. Alsubaei, Sajjan G. Shiva

PAGE 950 – 962

Paper 111: Authorship Attribution on Kannada Text using Bi-Directional LSTM Technique

Authors: Chandrika C P, Jagadish S Kallimani

PAGE 963 – 971

Paper 112: Flood Prediction using Deep Learning Models

Authors: Muhammad Hafizi Mohd Ali, Siti Azirah Asmai, Z. Zainal Abidin, Zuraida Abal Abas, Nurul A. Emran

PAGE 972 – 981

Paper 113: Recognition Method of Dim and Small Targets in SAR Images based on Machine Vision

Authors: Qin Dong

PAGE 982 – 990

Paper 114: Information Classification Algorithm based on Project-based Learning Data-driven and Stochastic Grid

Authors: Xiaomei Qin, Wenlan Zhang

PAGE 991 – 1000

Paper 115: Swine flu Detection and Location using Machine Learning Techniques and GIS

Authors: P. Nagaraj, A. V. Krishna Prasad, V. B. Narsimha, B. Sujatha

PAGE 1001 – 1009

**Paper 116: Taxation Transformation under the Influence of Industry 4.0**

*Authors: Pavel Victorovich Stroev, Rafael Valiakhmetovich Fattakhov, Olga Vladimirovna Pivovarova, Sergey Leonidovich Orlov, Alena Stanislavovna Advokatova*

**PAGE 1010 – 1015**

**Paper 117: Attractiveness of the Megaproject Labor Market for Metropolitan Residents in the Context of Digitalization and the Long-Lasting COVID-19 Pandemic**

*Authors: Mikhail Vinichenko, Sergey Barkov, Aleksander Oseev, Sergey Makushkin, Larisa Amozova*

**PAGE 1016 – 1021**

**Paper 118: Generation and Assessment of Intellectual and Informational Capital as a Foundation for Corporations' Digital Innovations in the "Open Innovation" System**

*Authors: Viktoriya Valeryevna Manuylenko, Galina Alexandrovna Ermakova, Natalia Vladimirovna Gryzunova, Mariya Nikolaevna Koniagina, Alexander Vladimirovich Milenkov, Liubov Alexandrovna Setchenkova, Irina Ivanovna Ochkolda*

**PAGE 1022 – 1032**

**Paper 119: An Algorithm for Providing Adaptive Behavior to Humanoid Robot in Oral Assessment**

*Authors: Dalia khairy, Salem Alkhalaf, M. F. Areed, Mohamed A. Amasha, Rania A. Abougalala*

**PAGE 1033 – 1039**

**Paper 120: Classifiers Combination for Efficient Masked Face Recognition**

*Authors: Kebir Marwa, Ouni Kais*

**PAGE 1040 – 1049**

**Paper 121: Human Position and Object Motion based Spatio-Temporal Analysis for the Recognition of Human Shopping Actions**

*Authors: Nethravathi P. S, Karuna Pandith, Manjula Sanjay Kofi, Rajermani Thinakaran, Sumathi Pawar*

**PAGE 1050 – 1056**



# ModER: Graph-based Unsupervised Entity Resolution using Composite Modularity Optimization and Locality Sensitive Hashing

Islam Akef Ebeid, John R. Talburt, Nicholas Kofi Akortia Hagan, Md Abdus Salam Siddique  
Department of Information Science  
University of Arkansas at Little Rock  
Little Rock, Arkansas

**Abstract**—Entity resolution describes techniques used to identify documents or records that might not be duplicated; nevertheless, they might refer to the same entity. Here we study the problem of unsupervised entity resolution. Current methods rely on human input by setting multiple thresholds prior to execution. Some methods also rely on computationally expensive similarity metrics and might not be practical for big data. Hence, we focus on providing a solution, namely ModER, capable of quickly identifying entity profiles in ambiguous datasets using a graph-based approach that does not require setting a matching threshold. Our framework exploits the transitivity property of approximate string matching across multiple documents or records. We build on our previous work in graph-based unsupervised entity resolution, namely the Data Washing Machine (DWM) and the Graph-based Data Washing Machine (GDWM). We provide an extensive evaluation of a synthetic data set. We also benchmark our proposed framework using state-of-the-art methods in unsupervised entity resolution. Furthermore, we discuss the implications of the results and how it contributes to the literature.

**Keywords**—Entity resolution; data curation; database; graph theory; natural language processing

## I. INTRODUCTION

Entity resolution is critical in data cleaning, curation, and integration [1]. It also refers to finding duplicate records within the same table, across various tables, or multiple databases that might refer to the same entity. Traditional and rule-based entity resolution relies heavily on human input to guide the entity matching process using predefined rules. Defining those rules depends on handcrafting simple lexical, semantic, and syntactic conditions for matching records based on attribute similarity, such as in [2] and in [3]. However, moving toward automating entity resolution for data cleaning, curation, and integration has become a sought-after goal in many domains. Thus, unsupervised entity resolution methods have increased.

Nevertheless, unsupervised approaches suffer from higher inaccuracies than other methods due to relying solely on approximate string matching. Approximate string matching algorithms can quantify the similarity between strings based on character or token frequency and location [4]. String similarity metrics can vary in granularity from character-based to context-based. For example, character-based approaches such as Levenshtein's Edit Distance [5], Affine Gap Distance [6], Smith-Waterman Distance [7], Jaro Distance [8], or n-gram based algorithms [9] are better suited for data where the order

of the tokens matter in identifying unique entities [4]. On the other hand, token-based approaches such as Overlap, Cosine, Dice, Monge-Elkan [10], and Jaccard [11] rely on tokenizing the text into a finite set and then comparing the intersection and union between the sets, which makes them better suited for data characterized by typographical errors. The TF-IDF algorithm [12] is another type of string similarity metric, which is more context-based and depends on token frequencies in a corpus.

Unsupervised entity resolution methods typically follow an automated processing pipeline that consists of preprocessing, blocking, matching, clustering, profiling, and canonicalization. Preprocessing refers to multiple steps that involve merging and parsing data files, tokenizing, and normalizing the unstandardized documents. Blocking is the strategy used to mitigate the quadratic complexity of pairwise comparisons in unsupervised entity resolution. That strategy relies on quick and dirty techniques that divide the preprocessed unstandardized references into chunks or blocks, avoiding string matching across the whole dataset. As a result, each block can be processed separately, where pairwise string similarity can be applied with less computational cost.

Generally, unsupervised entity resolution systems resort to matching threshold setting and end the entity matching process at that stage, such as in [13]. Other systems further expand the pipeline to identify entity profiles generalizing the entity matching output to more than two entity clusters. The clustering process aims to resolve conflicts in pairwise matching and find records that indirectly match. Those conflicts typically occur due to the reliance on frequency-based blocking [1]. Thus, to increase automation, reduce the amount of human input, and increase efficiency in the unsupervised entity resolution process, we aim to reduce the number of input parameters needed and to step away from direct approaches in approximate string matching. We introduce a graph-based approach to entity profiling in unsupervised entity resolution systems that leverage graph clustering algorithms' maturity and autonomy.

More specifically, we address the following challenges in graph-based unsupervised entity resolution systems represented by [13] and iterative self-assessing systems represented by [1]:

- The processing pipeline in iterative self-assessing systems might need to be applied multiple times due to the low accuracy of relying on approximate string

matching alone without rules as in traditional methods. That is clear in approaches such as the Data Washing Machine (DWM) [1].

- Using approximate string-matching similarity measures that rely on heuristics, though providing robust results, sometimes undermines processing speed. Moreover, those algorithms are exhaustive with quadratic time complexity running time, such as in [14]. In addition, though [13] introduced a learned similarity function, the algorithm is relatively expensive and does not provide an entity profiling capability.
- Setting matching thresholds as in [13], and cluster quality thresholds as in [1] might be problematic for the user's perspective. Interpreting those thresholds depends on the fed data, and the user might not have a baseline reference to compare.

#### A. Contribution

Here we developed a solution that relies on exploiting the transitivity property characterized by the output of the matching process, allowing us to recast the matched documents as a graph of weighted edges. We expand on the work that our group has done, mainly the Graph-based Data Washing Machine (GDWM) [15] and the original Data Washing Machine (DWM) [1]. We address the previously mentioned problems as follows:

- ModER avoids the extra computational cost of iterating multiple times over the processing pipeline to maximize a similarity threshold and optimize a cluster quality threshold [1]. First, the recast graph is divided into smaller subgraphs using a connected component detection algorithm exploiting the transitivity property of pairwise matching. Second, a document-word bipartite graph is formed where an initial modularity optimization runs to initialize cluster memberships of document nodes on the block level subgraph. Finally, a conditional greedy modularity maximization algorithm further breaks down the detected clusters.
- Instead of costly computing token-based similarity measures, the weighted edges representing the approximated similarity between every two documents on the block level are estimated using a Locality Sensitive Hashing scheme [16]. The similarity weight is approximated using a MinHash Jaccard estimator [17]. In addition, we exploit the fact that documents or records almost always include highly discriminative terms representing a fingerprint for each document.
- Instead of relying on the user to set similarity matching and cluster quality thresholds, we overcome the need to set algorithm-related thresholds by using Modularity as an optimized cluster quality metric guiding the matching and linking processes. The only parameters that the user needs to input are data-related: the percentile of blocking words, the percentile of stop words, and the percentile of discriminative words. That allows the user to study the data before running our framework statistically. The user can then provide those parameters as a function of the statistical analysis of the dataset.

- To our knowledge, Modularity based graph clustering has not been adapted before to the problem of unsupervised entity resolution. Therefore, we make our code publicly available as a git repository through the link under the directory ModER<sup>1</sup>.

## II. RELATED WORK

There is a large number of entity resolution systems in the literature in general targeting many problems such as ZeroER [18], DITTO [19] and Swoosh [20]. In this literature review we focus specifically on papers that are within the scope of graph-based unsupervised entity resolution. Despite their sparsity in the literature, graph-based methods and algorithms have been adapted before to entity resolution.

### A. Token-based Graph Entity Resolution

In token-based graph entity resolution, the goal is to construct a bipartite undirected graph of token nodes and record nodes and cluster the record nodes into unique entities using methods such as SimRank [21]. In [22], the authors introduced a graph-based entity resolution model. The model transformed the input data set into a graph of unique tokens where connectivity reflects the co-appearance of tokens in references. The graph was clustered using a weight-based algorithm that considered three types of vertices: exemplar, core, and support vertices. The algorithm then constructed  $r$  radius maximal subgraphs from the original token graph to discover clusters related to unique entities. Token-based methods, however, are computationally expensive and memory intensive due to the lack of an integrated blocking strategy.

### B. Record-Record Similarity-Based Graph Entity Resolution

Record-record similarity graphs link structured unstandardized references in a weighted undirected graph where the nodes represent unique records. The connectivity represents the degree of similarity between individual references. That approach of constructing a record graph allows to directly utilize a whole set of graph clustering algorithms that graph theory and network science researchers have already developed. While [23] applied a graph clustering algorithm to optimize minimal cliques in the graph. The algorithms approximated the NP-hard graph clique problem through pruning. Moreover, [24] developed the FAMER framework to combine multi-source data using blocking, matching, and clustering schemes. The framework modeled the merged data as a similarity-record graph and then leveraged graph clustering techniques to resolve the entities.

Other work has leveraged the graph's structure instead of just the weights between records. In [25], the authors proposed three algorithms to cluster the similarity graph based on structure rather than edge weights. They argue that graph-based transitive closure, such as in [26], produces high recall but low precision because the graph's structure is not considered during clustering. They justified using maximal clique algorithms to leverage the graph's structure, which increases precision. There are also centrality and node importance-based methods where the edge weights are not considered, and node scores are propagated, such as in [27]. In addition, the authors introduced

<sup>1</sup><https://bitbucket.org/oysterer/dwm-graph/src/master/ModER/>

the notion of a node resistance score in a co-authorship graph to model entity similarity. Node resistance can be considered a PageRank score [28], where a random walker computes the probability of getting from the source node to the target node iteratively until convergence. Also, in [29], the authors introduced a graph-based model that linked two graph datasets by aggregating similarity scores from neighboring record nodes. However, record-record similarity methods are complicated and require extensive graph theory knowledge to tune the adapted methods.

### C. Hybrid Graph-based Entity Resolution

Hybrid methods that combine token-based bipartite graphs and similarity-based record-record graphs have been investigated in [13] and [30]. The authors proposed an algorithm that combines text similarity with a graph-based algorithm. They first partition the data into a bipartite graph of record pair nodes and frequent term nodes to learn a similarity score of the record pair nodes. Then, the result was used to construct a record-record graph and used to power the CliqueRank algorithm, which runs on the blocks of records identified by the first part, known as the ITER algorithm. The probability of a matching pair of records is then updated iteratively. The authors combined two distinct methods: the random walk-based approach and the graph clustering-based approach. However, the authors used the graph approach to match pairs of records without introducing any clustering approach that would resolve the entity profiles.

The following section describes the framework, method, and algorithm used in ModER.

## III. PRELIMINARIES

### A. Problem Definition

Let us assume that we have a collection of merged documents in one file where every single document has a unique identifier and a reference body. More formally, the set of merged documents are  $D$  consisting of tuples  $d_i = (u_i, r_i)$  where  $u_i \in U$  is a unique identifier that is either provided in the input file or is automatically generated by ModER and  $r_i \in R$  is the reference body. Let us also assume that there exists a latent variable  $\tau$  representing the underlying hidden unique entity profiles in the data file pointing to the probability  $P(\tau) = \sum_{i=0}^N P(d_i \in \tau)$  where  $N$  is the number of documents in the file. Hence  $|\tau| \leq N$ . That reformulates the problem as an estimation of  $P(d_i \in \tau)$  for each document  $d_i$ .

### B. Graph Formulation and Modularity

Consider a set of document unique identifiers  $u_i$  and their tokenized unique reference bodies  $r_i$ . Consider two ways of remodeling the input corpus as a graph. First a document-document graph  $G = (V, E)$  where each vertex/node  $v \in V$  represents a unique document  $u_i \in U$  where  $u \equiv v$ . While an edge  $e \in E$  where  $E \subseteq V \times V$  represents whether two records  $e_i = (r_i, r_j)$  are matched, and an edge weight  $w_e \in W(E) : E \rightarrow \mathbb{R}$  represents the normalized similarity between the two nodes. A graph  $G$  could be represented as an adjacency matrix  $A$  of size  $|V| \times |V|$ , where each cell in the matrix contains either a 1 if an edge exists between two nodes or 0 if an edge does not exist. Each cell value containing 1 could

be multiplied by the edge weight to represent a weighted adjacency matrix. We also define a set of clusters  $Q \subseteq P(V)$  where  $Q$  elements are a subset of  $V$  and  $P$  is a partition of  $V$ . The clustered graph conventionally can be seen as a graph of subgraphs where each meta-vertex represents each subgraph of vertices or records as follows:

$$V' = Q \quad (1)$$

$$E' = (Q_i, Q_j) : \exists (v_i, v_j) : v_i \in Q_i, v_j \in Q_j, (v_i, v_j) \in E \quad (2)$$

Second, the corpus of input data could be modeled as a bipartite graph  $G = (V, Y, E)$  with two types of nodes  $V$  and  $Y$  where an edge  $e \in E$   $E \subseteq V \times Y$  can only exist between two nodes of different types. In our case, the first type of node  $v \in V$  represents a unique document  $u_i \in U$  where  $u \equiv v$  and the second type of nodes  $y \in Y$  represents a unique token  $t_i \in T$  where  $t \equiv y$ . Edges can exist between a document node and a token node if the token exists in the document reference body. We also define the notion of node membership  $q_i$  where a node  $n_i$  can only be a member of one cluster  $q_i$ . Finding the best suitable cluster membership for a node  $n_i$  could be achieved through optimizing cluster quality heuristics such as Modularity [31] or Conductance [32]. Modularity quantifies the cluster quality in a graph by comparing the edge density in each cluster to a randomly rewired hypothetical network. Recall the notions defined in equation 1 and 2. Modularity can then be conceived as:

$$M = \frac{1}{2m} \sum_{i, j} \left[ A_{i, j} - \frac{k_i k_j}{2m} \right] \partial(Q_i, Q_j) \quad (3)$$

Where  $m$  is the number of edges in the graph and  $k$  is the degree of a node. In addition  $A$  is a weighted adjacency matrix constructed from  $E'$ . And,  $i$  and  $j$  are indices for each unique record in the file, represented as a vertex in the graph as  $vt \in V'$ . And  $e' \in E'$   $E' \subseteq V' \times V'$  and  $e' \equiv A_{i, j}$ .

### C. Similarity and Transitivity

Consider that if record  $a$  matches record  $b$  and record  $b$  matches record  $c$ , then by transitivity, record  $a$  matches record  $c$ . The former definition of transitivity is the notion that binds our assumptions that lead us to create a graph from a set of matched documents. This assumption can only hold if the probability of record  $a$  matching record  $b$  and the probability of record  $b$  matching record  $c$  are high enough [33]. A high enough probability in approximate string matching is considered above 50% [34]. we interpret normalized approximate string similarity measures such as the Jaccard index and Levenstein ratio as matching probabilities. Hence, a Jaccard similarity between two documents below 50% is not accepted as a link between two records, which is crucial for the transitivity assumption to remain valid. Note also that the transitivity assumption is what allows us to form a document-document graph; otherwise, it does not make logical sense to apply a transitive closure algorithm on a formed graph if transitivity does not hold or, in other terms, if the edge weight between nodes representing the matching probability is less than 50%. That assumption could also be corroborated by interpreting similarity at 50% as extreme uncertainty of whether the two documents are similar instead of the intuition that a 10% similarity indicates

uncertainty. On the contrary, a 10% similarity holds more certainty that the two documents are dissimilar. Hence a 50% similarity is an appropriate baseline for interpreting approximate string similarity measures as matching probabilities.

#### IV. METHOD

In this section we extensively describe the proposed framework as shown in Fig. 1.

##### A. Merged Input Corpus

The input corpus is defined in one file. The file is merged from multiple data sources. The user manages this step where the only requirement is a single merged file. In the future, we intend to address having to merge multiple data sources as part of the framework developed here. The assumption is that each line in the file represents a document containing a reference body where the number of latent entity profiles is less than or equal to the number of documents. In addition, as mentioned in the preliminaries section, it is assumed that documents share at least one token so that the assumption of transitivity is not broken III-C. Note that the unique identifier  $u_i \in U$  is either provided in the input file or is automatically generated by ModER.

##### B. Preprocessing

The parsing process includes normalization, cleaning, tokenization, and filtering and is central to the framework. Tokenization is preceded by the normalization and cleaning step, where the text from the reference bodies of the documents is cleaned from particular characters and converted to lowercase. If a unique character appears in the token's middle, it is removed, and the entire string is compressed. We use the standard approach to tokenizing text in English, splitting the text based on spaces between tokens. The parsed data are then loaded into memory, and unique token frequencies and length dictionaries are computed. Stop words are also removed if their frequencies exceed a parameter sigma  $\sigma$ . More formally, as referred to before in the preliminaries Section III-B, the corpus  $D$  consists of tuples  $d_i = (u_i, r_i)$  where  $u_i \in U$  and  $r_i = w_i \in W$  contains a unique distinct set of tokens  $T$  with different counts  $C$  given that  $t_i$  is a distinct unique token where  $t_i \in T \subseteq w_i$  and  $c_i$  is the corresponding count of each unique token  $t_i$  where  $c_i \in C$  and  $C = |t_i \in T|$ . In addition, a token length dictionary is computed where  $l_i \in L$  and  $L = \text{len}(t_i \in T)$ . First each document  $d_i \in D$  is processed using  $\sigma$  to filter out tokens  $t_i \in r_i$  with frequency  $c_i \in C$  and  $C = |t_i \in T|$  above  $\sigma$  to filter out stop words.

##### C. Blocking

The blocking process aims at reducing the potential number of pairwise matching across a data file as much as possible. We use a frequency-based blocking algorithm that assumes that two records that refer to the same entity share at least one token. We adapted the frequency-based blocking technique presented in the DWM [1]. The blocking method relies on the parameter beta  $\beta$ . For each reference token  $t_i \in r_i$  with frequency  $c_i \in C$  and  $C = |t_i \in T|$  below  $\beta$  and above 2 is considered a blocking token  $t_{Bi} \in T$ . Blocking tokens are identified for each reference in a list  $L$  where the filtered

records are repeated. The list is then grouped by blocking tokens regardless of reference. Each block includes all the references where the same blocking token appeared, and the number of blocks was equivalent to the number of unique blocking tokens in the dataset. This process is formalized in Algorithm 1.

---

**Algorithm 1:** Blocking

---

```
Input :  $C(T)$ ,  $R$ ,  $\beta$ ,  $\sigma$ : unique token frequencies,  
record set  
Result :  $B$ : list of blocks  
 $B \leftarrow \text{list}$   
for  $r \in R$  do  
  if  $c_i \in C$  for each  $t_{ij} \in r$  where  $t_i \in T \geq \sigma$   
  then  
     $r \leftarrow \text{remove } t_{ij} \text{ from } r$   
  end  
  if  $c_i \in C$  for each  $t_{ij} \in r$  where  $t_i \in T \leq \beta$   
  then  
     $L \leftarrow (t_{ij}, r)$   
  end  
end  
 $L_s \leftarrow \text{sort}(L, t_i \in T)$   
 $T_u \leftarrow \text{unique}(t_{ij} \in L_s)$   
for  $t_i \in T_u$  do  
  for  $l_s \in L_s$  do  
    if  $t_i = t_i \in l_s$  then  
       $b_i \leftarrow r \in l_s$   
       $B \leftarrow B + b_i$   
    end  
  end  
end  
return  $B$ 
```

---

##### D. Fast Matching using Locality-Sensitive Hashing

The goal here is to provide a quick, fast, and multilevel way of grouping documents that might belong to the same latent unique entity. However, matching every document in the corpus would result in an algorithm that runs as  $O(N^2)$  in time. So, our efforts are concentrated on reducing the number of possible matching operations through 3 steps. First, after the preprocessing phase, we apply a Locality Sensitive Hashing (LSH) [16] algorithm to allow for a quick approximation of a similarity function such as Levenshtein ratio, Cosine distance, or Jaccard index. Even at the block level, computing the former similarity metrics can be expensive for larger files. LSH aims at using a hashing algorithm to approximate a similarity function such as Jaccard index [11]. Jaccard similarity involves computing the set intersection and union across tokenized words between the two documents being compared. Computing both a union and an intersection could be computationally expensive, misaligning with our overarching goal of reducing string-wise comparisons as much as possible. An LSH algorithm operates on a metric, a threshold, an approximation factor, and a set of probabilities. The goal is to design a hash function that approximates a metric by optimizing a threshold. The approximation is achieved by making sure that the set of probabilities holds.

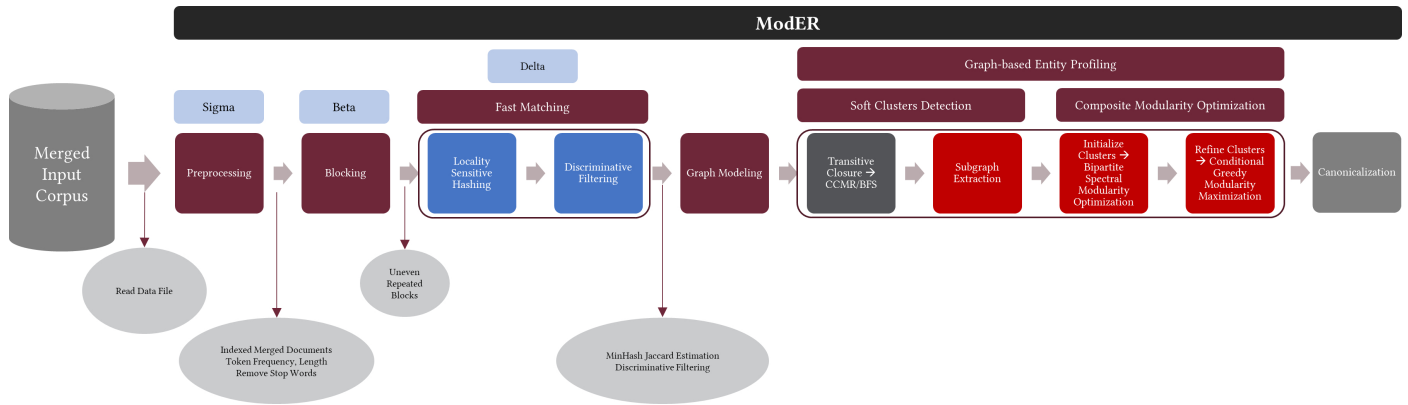


Fig. 1. An Overview of the ModER Framework.

Here we adapt the MinHash algorithm [17] to estimate the similarity between records in the same block. The MinHash algorithm is designed to approximate the Jaccard index between two sets. Sets are columns in a matrix where each row represents an element from the union of the two sets. A cell value of 1 represents the existence of that element in one of the two sets. The critical observation is that the Jaccard similarity could be approximated by counting the number of collisions between the two sets after hashing each element in each set. So counting the number of collisions of hashes between the two sets and normalizing them by the total number of hashes approximates the Jaccard similarity [35] [16]. That technique reduces the number of operations needed to compare two documents from  $O(N^2)$  to  $O(N)$  in time in best cases and  $O(N \log(N))$  in worst cases. we adapt MinHash by precomputing the approximating hash functions over the entire corpus. We provide a simple implementation of the MinHash algorithm in our code; however, while running experiments, we used the highly optimized implementation introduced by [36]. Each word is hashed for each document in the corpus using a 32-bit SHA algorithm [37]. The hashed unique tokens represented in each document are permuted and then drawn randomly from each document. A signature is then computed for each unique word in the documents, and the minimum signature is chosen to represent that word in the documents. The Jaccard index is then estimated linearly by counting the number of similar signatures in the same position over the total number of signatures appearing in the two documents according to the proof presented initially in [17]. We did not formalize MinHash here as formalizations of the algorithm are widely available.

The second part of our matching scheme, also done on the block level, is that we do not take the estimated Jaccard similarity at face value. We first filter the documents and extract what we call discriminate tokens. Those are tokens longer in length than a parameter delta  $\delta$ . Discriminate tokens represent tokens that, by looking at them, you can quickly determine whether two documents are similar. Those tokens might be social security numbers, credit card numbers, long street names, long last names, scientific names, product numbers, or models. Those tokens are usually highly discriminative in determining whether two documents are similar. Hence before estimating the Jaccard index on the block level, we check whether the two documents have tokens in common that are longer than

delta and their Levenshtein distance is less than 2. We consider 2 to be the threshold that defines a typo. The strength of our framework lies in the fact that we do not use a similarity threshold to match documents on the block level. Instead, we allow the data to automatically match the documents based on the characteristics represented in the parameters derived from the token frequencies and length. Hence, on the block level, we link documents with a similarity above 0 without any threshold setting. That process is described semi-formally in the pseudo-code presented in Algorithm 2.

---

**Algorithm 2:** Pairwise Matching using MinHash

---

**Input :**  $B, F$ : list of computed blocks, unique token lengths  
**Result :**  $E$ : linked weighted pairs  
 $E \leftarrow list$   
**for**  $b \in B$  **do**  
    **for**  $r1 \in b$  **do**  
        **for**  $r2 \in b$  **do**  
            **if**  $r1 < r2$  **then**  
                 $d1 \leftarrow f1 \in F$   
                 $d2 \leftarrow f2 \in F$   
                **if**  $d1 = d2$  **or**  $levenshtien(d1, d2) \leq 2$   
                    **then**  
                         $s \leftarrow 1.0$   
                    **end**  
                **else**  
                     $s \leftarrow minHash(r1, r2)$   
                **end**  
                **if**  $s > 0.0$  **then**  
                     $E.append((r1, r2, s))$   
                **end**  
            **end**  
        **end**  
    **end**  
**end**  
**return**  $B$

---

**E. Graph Modeling**

The unordered list of matched records can be considered an edge list or an adjacency matrix representing a graph of

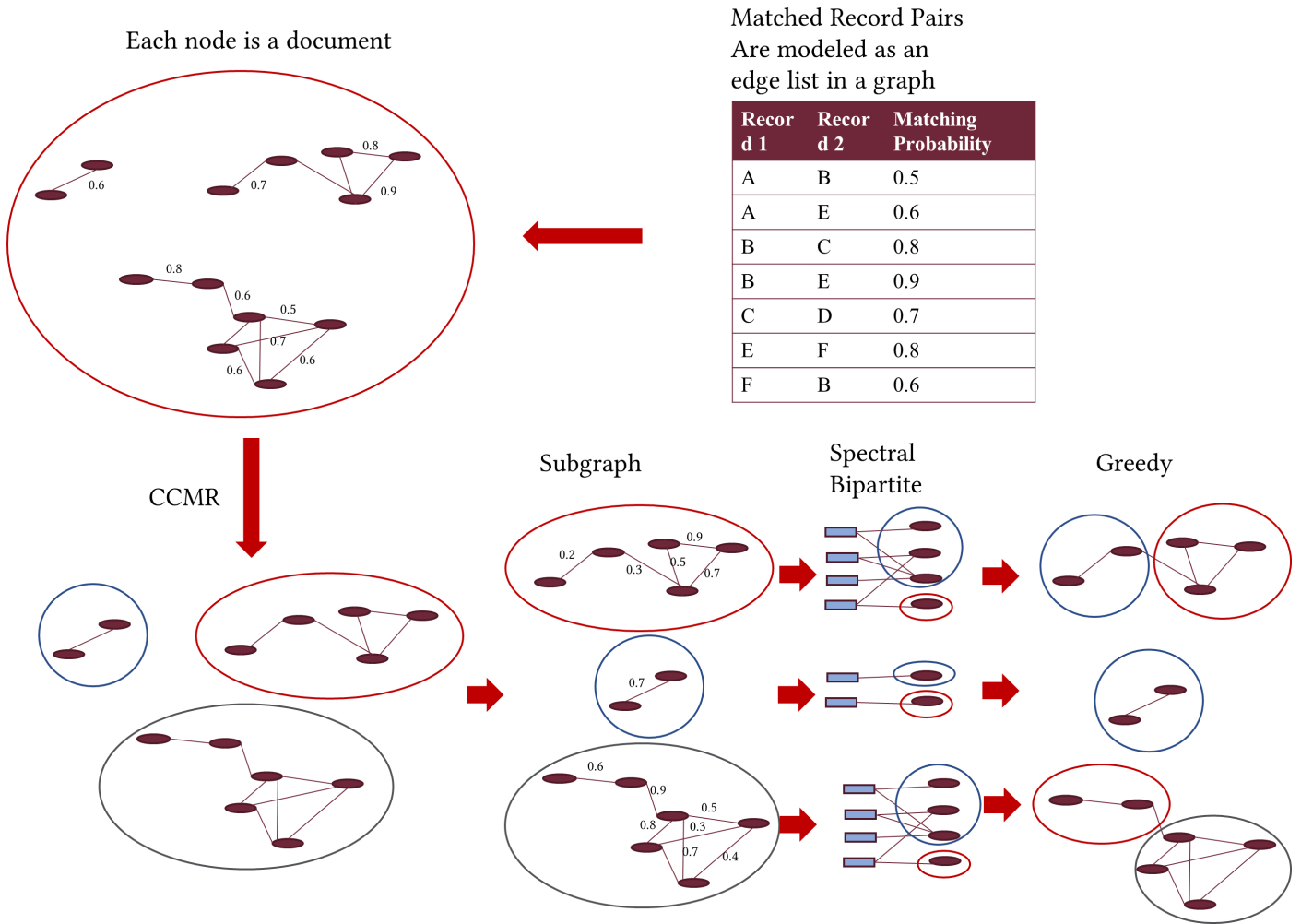


Fig. 2. Graph Modeling. First, the Matched Records are Modeled as One Large Weighted Graph with Expected Disconnected Components Due to the Transitivity Property. Second, Connected Component Detection Algorithms Output the Nodes in each Connected Component. Third, each set of Nodes is then Modeled as a Subgraph.

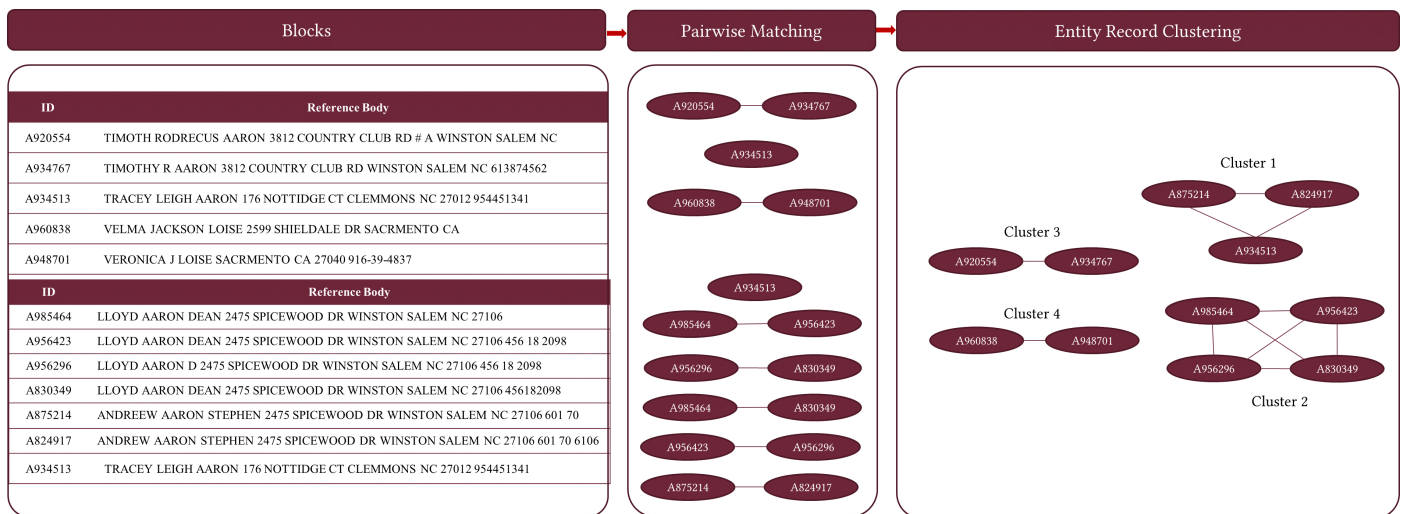


Fig. 3. The Typical Unsupervised Entity Resolution Approach Starts with Frequency-Based Blocking, Pairwise Matching, and Entity Clustering.

records. The blocking algorithm considers two similar records having a high probability of representing the same entity if



they have a minimum of one token in common. That increases the number of records representing one entity, as shown in Fig. 3. As a result, a record may appear in more than one entity, making clustering more expensive and the underlying graph of matched records more complex. In addition, the matching process also implies that one record might be similar in an indirect way to another record by transitivity, provided that the matching probability is above 50%. The output of algorithm two is modeled as a graph, as mentioned in the preliminaries in Section III-B and as seen in Fig. 2. A subgraph is then extracted on the block level.

#### F. Graph-based Entity Profiling

1) *Transitive Closure*: In transitive closure, the aim is to find the set  $C$  of sets  $V_i$  where each set represents a strongly connected component in the graph. A strongly connected component is a set of nodes where each node has at least one link to another node. Strongly connected components are separated by unlinked nodes. We also refer to strongly connected components as soft clusters. Here in ModER, one of two approaches could be used. First, an algorithm named CCMR was introduced by [26] and used and reimplemented in the original DWM [1] as a single-step clustering method. The algorithm starts by creating star subgraphs from the matched pair list. That is done by simply grouping the pair list by the smallest node. Then, the algorithm iteratively checks whether its vertices are assigned to subgraph components where the center of the subgraph component is the smallest vertex in its first-order neighborhood, relying on the MapReduce framework for scalability. The DWM provides an efficient implementation of the algorithm without relying on MapReduce. The algorithm is formalized and described thoroughly in [26] and [1]. The second approach that could be used is a simple breadth-first search algorithm [38] to extract the connected components or soft clusters. We provide implementation and describe both algorithms semi-formally in Algorithms 3 and 4.

2) *Subgraph Extraction*: The subgraph extraction process aims to avoid recomputing edge weights that have been computed before during the fast-matching process. A subgraph is simply the set of soft clusters that have been computed using the transitive closure process. Each soft cluster is represented by a set of unique nodes or documents. Each document or node can appear in only one soft cluster. To extract a subgraph from the larger graph, we find the edges where all the nodes in the current soft cluster are represented exclusively. Subgraph extraction is formalized in Algorithm 5.

3) *Composite Modularity Optimization*: In the Composite Modularity Optimization step, the goal is to initialize the cluster membership of each node in the extracted subgraph based on token memberships in each document. Next is to refine the clusters discovered in the initial clustering to more precise memberships. In the following two subsections, we outline how we implement that process.

a) *Bipartite Spectral Modularity Optimization*: This step exploits the relationship between documents and unique tokens across all documents. First, we project the unique document identifiers and unique tokens across all documents as a bipartite graph similar to Fig. 4.

---

#### Algorithm 3: Transitive Closure using Adapted CCMR

---

```
Input :  $S, \mu$ : pairs of matched records and their  
         computed similarity scores, similarity threshold  
Result :  $P$ : list of soft clusters as pairs  
         indexed by the least record in the cluster as  
         the first element of the pair  
 $P \leftarrow$  list of soft clusters  
 $R_P \leftarrow$  initialize list of tuples  
for  $s \in S$  do  
    if  $s_i \geq \mu$  then  
         $R_P \leftarrow$  append  $s_i$   
    end  
end  
 $R_P \leftarrow$  sort by first element in pair  
while no convergence do  
    for  $(r_i, r_j) \in R_P$  do  
         $P \leftarrow$   
            append pair belonging to connected component  
            when assuring that all record  
            nodes belong to the connected component  
            with the smallest record id  
            node at the center in their neighborhood;  
    end  
end  
return  $P$ 
```

---

---

#### Algorithm 4: Connected Components Breadth First Search

---

```
Input :  $G = (V, E)$ : graph  
Result :  $V'$ : set of nodes as component  
 $seen \leftarrow$  set  
 $components \leftarrow$  list  
for  $v \in V$  do  
    if  $v! \in seen$  then  
         $visited \leftarrow$  set  
         $queue \leftarrow$  list  
         $visited.add(v)$   
         $queue.add(v)$   
        while  $queue$  do  
             $dequeued \leftarrow queue.pop()$   
             $neighbors \leftarrow G.getNeighbors(dequeued)$   
            for  $n \in neighbors$  do  
                if  $n! \in visited$  then  
                     $visited.add(n)$   
                     $queue.append(neighbors)$   
                end  
            end  
        end  
         $seen.add(visited)$   
         $components.append(visited)$   
    end  
end  
return  $components$ 
```

---

A920554	1	TIMOTHY	RODRECUS	AARON	3812 COUNTRY CLUB RD # A	WINSTON SALEM	NC	27104	613-87-4562
A934767	2	TIMOTHY	RODRECUS	AARON	3812 COUNTRY CLUB RD # A	WINSTON SALEM	NC	27104	613-87-4562
A934513	3	TRACEY	LEIGH	AARON	176 NOTTIDGE CT	CLEMMONS	NC	27012	954451341
A960838	4	VELMA	JACKSON	AARON	2599 SHIELDS DR	WINSTON SALEM	NC	27107	559-64-4215
A948701	5	VERONICA	OLIVIA	AARON	3401 LOCHURST CT	PFAFFTOWN	NC	27040	335-39-4837

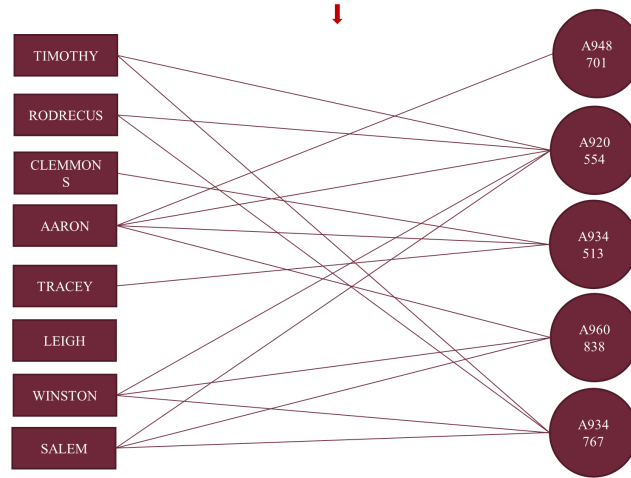


Fig. 4. A Token-Record Bipartite Graph.

**Algorithm 5: Subgraph Extraction**

```

Input :  $V', A$ : component, adjacency
Result :  $G' = (V', E')$ : graph
 $E' \leftarrow list$ 
for  $v \in V'$  do
     $neighbors \leftarrow G.getNeighbors(v)$ 
    for  $nn \in neighbors$  do
        if  $nn \in V'$  then
             $E'.append((v, nn, A[v][nn]))$ 
        end
    end
end
return  $(V', E')$ 
    
```

Recall equation 3; we defined the Modularity heuristic  $M$  as the sum of the difference between the edges of the current graph versus a null model graph where edges are rewired randomly within clusters. Here we approach Modularity optimization on the bipartite network of documents and unique tokens or words using a spectral approach or, in other words using matrix form. The goal is to optimize Modularity across both types of nodes in the bipartite graph and extract only the optimized cluster memberships of document nodes. First, a clustered index matrix  $S$  is defined where the rows are the nodes in the graph, and the columns are the number of clusters in the graph. Initializing this matrix at the beginning means that each node has its cluster, so the number of rows equals the number of columns. The values in  $S$  are either 0 or 1, indicating node memberships. Recall from equation 3 defining Modularity, the term  $\frac{k_i k_j}{2n}$  defines the probability of an edge in the null model

and can be denoted by  $P_{i,j}$ . Hence the term  $[A_{i,j} - \frac{k_i k_j}{2n}]$  can be rewritten as  $[A_{i,j} - P_{i,j}]$  and can be referred to as  $B_{i,j} = [A_{i,j} - P_{i,j}]$ . According to [39], the previous matrix form can be combined with equation 3 to redefine Modularity in matrix form as:

$$M = \frac{1}{2n} Tr(S^T B S) \tag{4}$$

Where  $n$  is the number of edges in the graph since a bipartite graph is naturally partitioned into a minimum of two initial clusters [39]. That naturally help us derive a definition of Spectral Bipartite Modularity that penalizes any random choices of edges if the nodes belong to the same cluster, thus the bipartite Modularity in a non-matrix form could be defined as:

$$M = \frac{1}{n} \sum_{i=1}^p \sum_{j=1}^q \left[ \widetilde{A}_{i,j} - \frac{k_i d_j}{n} \right] \partial(Q_i, Q_{j+p}) \tag{5}$$

Where  $\widetilde{A}$  is the bi-adjacency matrix.  $k_i$  is the degree of node  $i$  from the document nodes and  $d_j$  is the degree of node  $j$  from the word nodes.  $n$  is the total number of edges in the graph and  $Q_i$  is the partition of a document node  $i$  while  $Q_{j+p}$  is the partition of a word node  $j$  indexed as  $j = i + p$ . Thus equation 5 in matrix form becomes:

$$M = \frac{1}{2n} Tr(R^T \widetilde{B} T) \tag{6}$$

Where  $R$  is the cluster index matrix similar to  $S$  but for document nodes, while  $T$  is the cluster index matrix similar to  $S$  but for token or word nodes.  $\widetilde{B}$  is the difference between

the biadjacency matrix  $\tilde{A}$  and the reformulated bipartite null model  $\tilde{P}$  where  $\tilde{B} = [\tilde{A} - \tilde{P}]$ . The goal is to maximize Modularity in matrix form as in equation 6. The term  $\tilde{B}T$  in equation 6 could be rewritten as  $[\tilde{A}T - \tilde{P}T]$  indicating that the bipartite null model is calculated on the token nodes only. We refer to this case as  $\tilde{T}$ . Or it could be written as  $[\tilde{A}R^T - \tilde{P}R^T]$  indicating that the bipartite null model is calculated on the document nodes only. We refer to this case as  $\tilde{R}$ . In terms of summation over the matrix in an optimization scheme, these two cases can be rewritten as:

$$M = \frac{1}{n} \sum_{i=1}^p \left( \sum_{k=1}^c [R_{i,k} \tilde{T}_{i,k}] \right) \quad (7)$$

$$M = \frac{1}{n} \sum_{j=1}^q \left( \sum_{k=1}^c [\tilde{R}_{i,k} T_{i,k}] \right) \quad (8)$$

Where  $c$  is the maximum number of clusters in the graph, note that this is equivalent to hypothetically assigning document nodes to clusters of word tokens and vice versa, similar to traditional modularity optimization approaches such as Louvain [40] hence the delta modularity, in this case, was the summation of both delta modularities from both node types. Given the previous formulation, we maximize Modularity using a greedy approach described in Algorithm 6.

*b) Conditional Greedy Modularity Maximization:* This second step aims first to recast the documents as a document-document graph as in Fig. 5. Second to break down the soft clusters in a hierarchical manner by maximizing Modularity  $M$  using a conditional modularity maximization algorithm as seen in Fig. 2. That step is seen as further filtering. The assumption here is that clusters in a graph are often defined by the density of edges between a set of nodes. Hence, the number of connections or edges between clusters characterized by high Modularity is often meager. Modularity is defined in equation 3 as  $M$ . It measures the difference between the actual number of edges and an expected number of edges between nodes. An expected number of edges between nodes can be considered a random rewiring of the graph given the same nodes. Optimizing Modularity through maximization in a graph is a complex problem and is often tackled through various ways to reduce the number of comparisons between all nodes in a graph. During maximization, we use the shorthand equation provided in [40] to reduce the computational cost of computing delta Modularity of all nodes as shown in equation 9.

$$\Delta M = \left[ \frac{\sum_{in} + k_{i,in}}{2n} - \left( \frac{\sum_{tot} + k_i}{2n} \right)^2 \right] - \quad (9)$$

$$\left[ \frac{\sum_{in}}{2n} - \left( \frac{\sum_{tot}}{2n} \right)^2 - \left( \frac{k_i}{2n} \right)^2 \right] \quad (10)$$

Where  $M$  is Modularity,  $in$  are incident nodes to cluster,  $tot$  all nodes inside cluster,  $k$  is the degree of the node,  $n$  is the total number of edges. We also apply a condition that a node is only assigned to the maximum Modularity difference cluster if the delta modularity is positive and the matching probability between both document references is 100%. That

---

**Algorithm 6:** Spectral Bipartite Modularity Optimization

---

**Input :**  $V', D', W'$ : set of nodes in current soft cluster, document references, unique words  
**Result :** updated node labels  
 $edges \leftarrow list$   
 $p = length(V')$   
 $q = length(W'[V'])$   
 $c = p + q$   
 $rg = index(W'[V'])$   
 $gn = index(V')$   
 $A = zeroMatrix(p, q)$   
**for**  $v \in V'$  **do**  
    **for**  $w \in D'[v]$  **do**  
         $edges.append((v, w))$   
         $A[gn[v], rg[w]] = 1.0$   
    **end**  
**end**  
**for**  $i = 0$  **and**  $V'$  **and**  $i++$  **do**  
     $assignMembership(V'[i], i)$   
**end**  
**for**  $i = i$  **and**  $W'[V']$  **and**  $i++$  **do**  
     $assignMembership(W'[V'][i], i)$   
**end**  
 $ki = sum(A, 1), dj = sum(A, 0), kd = kidj$   
 $m = sum(ki), B = A - (kd/m)$   
 $T0 = initializeModularityMatrix(gn, V')$   
 $R0 = initializeModularityMatrix(rg, W'[V'])$   
 $minimumDeltaModularity = min(1/m, 0)$   
 $deltaModularity = 1, modPrevious = 0$   
**while**  
     $deltaModularity > minimumDeltaModularity$   
**do**  
     $Tp = T0^T . B$   
     $maximumModularityIndex = argMax(Tp^T)$   
     $R = zeroMatrix(q, c)$   
    **for**  $i, length(maximumModularityIndex)$  **do**  
         $R[i, maximumModularityIndex[i]] = 1$   
    **end**  
     $Rp = B . R$   
     $maximumModularityIndex = argMax(Rp)$   
     $T = zeroMatrix(p, c)$   
    **for**  $i, length(maximumModularityIndex)$  **do**  
         $R[i, maximumModularityIndex[i]] = 1$   
    **end**  
     $T0 = T$   
     $modCurrent = modPrevious$   
     $RtBT = T^T . B . R$   
     $sumMod = (1/m) * RtBT$   
     $modCurrent = sum(modCurrent)$   
     $deltaModularity =$   
         $modCurrent - modPrevious$   
**end**  
 $updateNodeMemberships(extractMemberships(T))$

---

Cluster ID	Reference ID	Reference Body
A824917	A985464	LLOYD AARON DEAN 2475 SPICEWOOD DR WINSTON SALEM NC 27106
A824917	A956423	LLOYD AARON DEAN 2475 SPICEWOOD DR WINSTON SALEM NC 27106 456 18 2098
A824917	A956296	LLOYD AARON D 2475 SPICEWOOD DR WINSTON SALEM NC 27106 456 18 2098
A824917	A830349	LLOYD AARON DEAN 2475 SPICEWOOD DR WINSTON SALEM NC 27106 456182098
A824917	A875214	ANDREEW AARON STEPHEN 2475 SPICEWOOD DR WINSTON SALEM NC 27106 601 70 6106
A824917	A824917	ANDREW AARON STEPHEN 2475 SPICEWOOD DR WINSTON SALEM NC 27106 601 70 6106

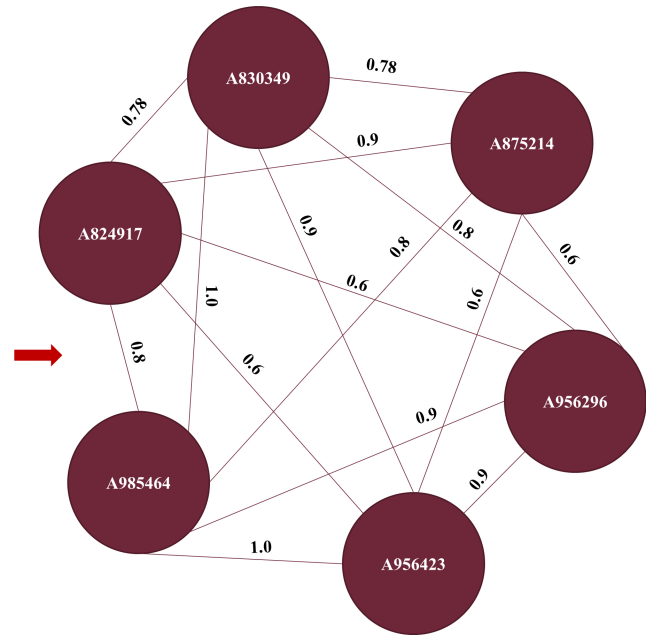


Fig. 5. Modeling a Data Set as a Record-Record Weight Graph based on the Similarity between the Records.

ensures that precision is not compromised by breaking precise clusters discovered in the first step. In addition, that is also to assure that discriminative terms are factored into the composite Modularity optimization method. Our Conditional Modularity Maximization algorithm is formalized in Algorithm 7.

**Algorithm 7:** Conditional Greedy Modularity Optimization

```

Input :  $G' = (V', E')$ : subgraph nodes
Result : updated node labels
 $M_{initial} \leftarrow computeModularity(V', E')$ 
for  $i \in V'$  do
     $iCom = G'.getMembership(i)$ 
     $neighbors = G'.getNeighbors(i)$ 
    for  $j \in neighbors$  do
         $jCom = G'.getMembership(j)$ 
         $weight \in G'.getEdgeWeight(i, j)$ 
         $\Delta M \leftarrow compute\ difference\ in\ Modularity\ M$ 
        according to Louvain's equation
         $L_M \leftarrow (j, \Delta M)$ 
    end
     $c_i \leftarrow \arg\ max(L_M)$ 
    find the largest cluster with the largest delta
    Modularity
     $\Delta M$ 
    and the corresponding weight
     $v2 \leftarrow v1$  assign the current node to the maximum
    cluster if edge weight is 1
     $updateMemberships(V')$ 
end
 $M_{final} \leftarrow computeModularity(V', E')$ 

```

G. Canonicalization

The clusters representing unique entity profiles are defined and persisted to the disk in this final step through a link index file. The link index file describes the final entity clusters of the unsupervised entity resolution framework as a list of ordered pairs where the first element of each pair represents the entity profile identifier to which the record is assigned. The entity profile identifier is simply the least recorded identifier in the cluster. The second element of each pair represents a member's unique record identifier.

V. EXPERIMENTS

A. Datasets

1) *Synthetic Datasets:* ModER is tested on a synthetic benchmark dataset. The reason lies in the fact that there is a lack of benchmark datasets for the task of entity resolution. Most entity resolution systems are evaluated against benchmark datasets [41] that are designed for an entity matching task and not an entity resolution task. Hence we use the synthetic dataset described in [42] as seen in Fig. 5 that is specifically designed for the task of entity resolution. This simulator-based data generator uses probabilistic approaches to generate coherent individual data for persons that do not exist except for S3 and S6, which both represents generated addresses and names of restaurants that do not exist and were introduced in [43]. The data fields are names, addresses, social security numbers, credit card numbers, and phone numbers mixed in several layout configurations. Some samples are labeled as mixed layout, meaning that each row might come with a different order of those attributes and might not be delimited. The standard label means that all the rows in the data file have the same order and attributes. The generator described in [44] used a probabilistic error model to inject various errors in the previously developed simulated dataset. For example, in this excerpt of a generated data file shown in Fig. 5, the first four records are almost

identical except for that record A956296 has a missing last name and the format of the phone numbers or whether they exist at all, all are errors injected and generated on purpose. In addition, the last two records are almost identical except for the first name, where an intentional error was introduced. A ground truth set recording the actual clusters of the simulated records is then sampled from the generated synthetic database. Next, the corresponding references are pulled from the generated synthetic database to create various sample files with different sizes, levels of quality, and layouts. Sizes of files can vary from 50 to 20K rows. For a more detailed description of the dataset please refer to [15].

2) **Benchmark Dataset: Abt-Buy** is an e-commerce benchmark dataset introduced and curated by [41]. Despite the lack of benchmark datasets that are designed specifically for the task of entity resolution rather than entity matching we used this dataset as a means to compare results with other systems. The dataset represents an entity matching task between products listed on 2 e-commerce websites<sup>2</sup>. The dataset appears as one data file containing attributes, product names, descriptions of products, manufacturer of the products, and price. The data set also is provided with a ground truth file containing a perfect matching between entities across both websites. The number of rows in this dataset is 1081 from the first source file and 1092 from the second data file.

3) **Benchmark Models: Data Washing Machine (DWM):** the Data Washing Machine (DWM) [1] implements a fully probabilistic, iterative, and token frequency-based approach. The DWM produces remarkable results on simulated datasets of various sizes, layouts, and qualities. The DWM is computational, iterative, and probabilistic; it follows a traditional data cleaning and merging pipeline, tokenization, blocking, matching, clustering, profiling, and canonicalization. The DWM relies on the idea of starting with a set of configuration parameters and then iteratively incrementing the parameters according to a self-administered evaluation of the quality of the clusters using an entropy-based metric. The DWM has been shown to provide robust results in large datasets. In addition, the current implementation is very modularized, allowing for room for improvement.

**Magellan:** Magellan, introduced in [45], is an end-to-end solution to entity resolution developed as a user-centric approach. It has been used and adopted widely. The model relies on providing guides that tell users what to do in entity matching scenarios. The framework also provides tools to cover the entire entity matching pipeline using a simple, approachable implementation.

## B. Evaluation Metrics

For evaluation we measured the precision and recall against the generated ground truth entity clusters. The ground truth is a list of each record and its membership cluster identifier. After canonicalization, the saved link index is grouped by the least record identifier in each cluster. Thus, all records belonging to the same cluster have the same record identifier as the first element in the link index pair. We then loop on each pair in the canonicalized link index and examine whether they belong together in the ground truth. Finally, we measure the following

statistics against the ground truth for each sample run as shown in Table I. That is also shown in the increased balanced accuracy [46], which is a valuable measure for problems such as entity resolution. Entity resolution is characterized by having a class imbalance as the number of matched pairs is usually way less than the number of unmatched pairs causing a very high number of true negatives.

## C. Results

We ran our model on an Intel Core *i7 – 4720HQ* CPU @ 2.60GHz and 32GB of RAM. We ran ModER and DWM first on the samples *S1* through *S18*, as described in Tables III and II. To determine optimal parameters for ModER relative to maximum F1-Scores, we ran each sample 10 times on incrementing beta, sigma, and delta and chose the parameters that gave the best F1-Scores. Please note that we do not set them directly as numbers when setting parameters beta  $\beta$ , sigma  $\sigma$ , and delta  $\delta$ . Instead, we set them using a percentile formula where the percentile of stop words is  $\sigma$ , the percentile of blocking words is  $\beta$ , and the percentile of word length is  $\delta$ . For running the DWM, we set the parameters to the equivalent in our model. For example, we assume that our baseline matching threshold  $\mu$  is 0.5 or 50% quantifying maximum uncertainty. Also, in the DWM runs, the quality epsilon  $\epsilon$  is tuned based on our previous work's data [15]. We also informed the setting of  $\beta$  and  $\sigma$  similar to what we did in the GDWM and set them to 6 and 7, respectively.

Table II shows the results from running the DWM on the 18 samples with equivalent set parameters. Even though it is challenging to compare both techniques due to their differences, it is helpful to compare both in relatively limited runs. Table V compares the final F1-Scores of both the DWM and ModER. On the other hand interpreting Table IV is tricky because more experimentation needs to be done to get a complete picture of the performance of the developed approach concerning what has already been done. However, ModER improves overall precision at the expense of recall from those results. Balancing precision and recall is challenging for most classifiers.

Composite Modularity Optimization as a technique for entity profiling is very efficient at detecting large clusters and breaking them down due to the reliance on a bipartite technique of modeling document nodes and unique words. That initialization is used in the second stage of Conditional Greedy Modularity maximization to refine the detected clusters further. Note that the conditional aspect of the greedy Modularity Maximization technique is intentionally injected into the algorithm to ensure that no nodes are assigned to new clusters if the delta modularity change is positive yet too little. That is also to ensure that we only see similarity at 100% or 1.0 as the most certain value indicating matching, and anything else below 100% is mere speculation and prediction and should be treated that way in other entity resolution systems.

In addition, when averaging all samples, as in Table V, ModER offers less F1-Scores performance than the GDWM and the DWM. Note that those averages were directly taken from [15]. Those values were for matching probabilities set at 70% and higher. That might not be fair because ModER does not rely on matching threshold settings. ModER could be seen as a quick and initial entity profiler before using other

<sup>2</sup>www.buy.com and www.rakuten.com

TABLE I. EVALUATION METRICS AND STATISTICS

Statistic	Symbol	Description
True Positives	TP	The number of record pairs that appeared together in the same cluster correctly
True Negatives	TN	The number of record pairs that did not appear in the same cluster correctly
False Positives	FP	The number of record pairs that appeared together in the same cluster falsely
False Negatives	FN	The number of record pairs that did not appear in the same cluster falsely
Precision	P	$TP / (TP + FP)$
Recall	R	$TP / (TP + FN)$
F1-Score	F1	$2 \times (P \times R) / (P + R)$
Balanced Accuracy	A	$TP \times (TN + FP) + TN \times (TP + FN) / (TP + FN) \times (TN + FP)$

TABLE II. RESULTS FROM RUNNING DWM ON EQUIVALENT PARAMETER SETTINGS

Sample	Precision	Recall	F1-Score
S1	76.47	96.3	85.25
S2	62.69	87.5	73.05
S3	41.15	95.54	57.52
S4	70.47	87.27	77.98
S5	71.84	87.94	79.08
S6	81.95	75.63	78.66
S7	73.57	86.65	79.58
S8	67.32	36.64	47.45
S9	64.22	17.48	27.48
S10	72.85	27.58	40.01
S11	70.69	26.19	38.22
S12	73.27	26.37	38.78
S13	76.64	83.32	79.84
S14	70.43	84.73	76.92
S15	68	82.04	74.36
S16	74.02	26.84	39.4
S17	70.59	24.39	36.25
S18	69.79	30.33	42.28

systems or rely on a highly unsupervised system that does not require the user to set a matching threshold. In addition to a system that does not rely on matching threshold settings, unlike most state-of-the-art systems, the results are comparable and tell something about the need for matching thresholds. When a user uses such a system to detect entity profiles in a file, they have no experience with what a matching threshold might mean. Hence in ModER, we only let the user set 3 explainable parameters using percentiles, assuming they have studied their data statistically before running any entity resolution system.

In Table VI, we benchmarked ModER using the Abt-Buy dataset introduced in [41]. In addition, Table VII provides insight into multiple runs of ModER on the Abt-Buy dataset with different parameter configurations. Table VI

reports running ModER with parameters that resulted from 10 times increasing parameters with the best F1-Score. As seen, ModER outperforms our previous system, DWM, in addition to Magellan. That is mainly due to the Composite Modularity Optimization algorithm efficiency. In addition, our reliance on discriminative words has favored ModER since most data contain a higher percentage of discriminative keywords. In Table VII, it appears that the number of single nodes varied widely as with the number of edges in the graph formed after matching. That is due to the threshold of the varying parameter. In addition, those initial parameter settings affect the initial Modularity widely regardless of the final Modularity. A Higher F1-Score was tied to lower  $\sigma$  levels, suggesting that filtering stop words are always beneficial before running any entity matching algorithm.



TABLE III. RESULTS FROM THE SYNTHETIC DATA SET EXPERIMENTS ON MODER. BETA IS THE PERCENTILE OF BLOCKING WORD FREQUENCIES ACROSS THE WHOLE FILE. SIGMA IS THE PERCENTILE OF STOP WORD FREQUENCY ACROSS THE WHOLE FILE. FURTHERMORE, DELTA IS THE PERCENTILE OF TOKEN LENGTHS ACROSS THE DATA FILE. BA IS THE BALANCED ACCURACY AND IS COMPUTED AS DESCRIBED IN TABLE I. FINALLY, THE FINAL MODULARITY IS MEASURED ON THE FULL DISCONNECTED GRAPH AFTER ASSIGNING THE NEW MEMBERSHIPS AFTER THE COMPOSITE MODULARITY OPTIMIZATION STEP.

Sample	Beta	Sigma	Delta	BA	Precision	Recall	F1	Modularity
S1	80%=3	99%=31.16	100%=17	98.15	100	96.3	98.11	76.74
S2	80%=3	100%=95	80%=9	93.73	95.45	87.5	91.3	66.76
S3	80%=2	95%=7	69%=9	94.19	83.19	88.39	85.71	96.13
S4	90%=5	95%=8	70%=9	86.82	91.35	73.64	81.54	61.21
S5	90%=5	95%=8	70%=9	87.55	93.78	75.1	83.41	62.11
S6	94%=6	99.99%=1972	40%=6	64.03	71.37	28.06	40.28	33.72
S7	93%=6	99%=36.5	95%=9	85.69	92.17	71.39	80.46	56.66
S8	95%=7	100%=400	50%=5	58.99	67.33	18.04	28.45	20.91
S9	95%=7	100%=322	60%=6	58.93	59.47	17.93	27.56	32.88
S10	80%=6	99%=32.5	60%=6	58.58	71.15	17.21	27.71	53.76
S11	80%=5	100%=1458	50%=6	56.25	76.39	12.51	21.5	50.36
S12	80%=4	100%=1859	30%=5	54.51	79.43	9.0	16.2	47.88
S13	90%=6	100%=1887	45%=5	81.05	68.65	62.13	65.23	42
S14	90%=7	100%=4738	90%=9	77.8	79.82	55.6	65.54	43.98
S15	90%=7	100%=9447	90%=9	76.81	81.82	53.62	64.79	46.34
S16	80%=5	100%=713	30%=5	57.01	77.0	14.05	23.76	51.84
S17	75%=4	100%=17.91	30%=5	54.11	78.92	8.23	14.9	50.52
S18	75%=4	100%=3405	30%=5	53.84	78.26	7.69	14.0	51.53

## VI. DISCUSSION

### A. Overall Effectiveness

Ideally, we need a better measure to quantify the balance between precision and recall, as seen in Fig. 6. The F1-score or the harmonic mean between precision and recall fails to differentiate between instances where recall or precision was very high and when they were balanced. A better entity resolution system always provides a balance between both. In that light, ModER appears to balance both precision and recall on S1, S2, S3, and S13. Despite their relative diversity, the common characteristic between those samples is relatively higher quality. That is, the difference between document references injected errors is not significant. They indicate that the first bipartite spectral Modularity optimization did all the work to detect entity profiles. The problem is that bipartite Modularity optimization is a memory-intensive optimization for more significant clusters even though its time complexity is at

$O(2N)$ , returning only two passes on the Modularity matrix to compute the difference. That points us to address this limitation in the future of balancing space and time complexities.

### B. The Effect on Modularity

Here we refer to the final Modularity as the Modularity computed on the final overall graph that has been projected before the entity profiling step. The final cluster memberships have been computed using our Composite Modularity Optimization approach.

In Fig. 7, modularity is more correlated with precision than recall. The more clusters are broken down during the entity profiling step, the higher the Modularity is. Note that Modularity is weakly correlated with the F1 score, meaning that higher modularity values do not necessarily mean higher F1 scores. Modularity is a comparison of edge densities between

TABLE IV. COMPARISON BETWEEN DWM AND MODER

Samples	DWM F1-Score	ModER F1-Score	Performance
S1	85.25	98.11	Improved
S2	73.05	91.3	Improved
S3	57.52	85.71	Improved
S4	77.98	81.54	Improved
S5	79.08	83.41	Improved
S6	78.66	40.28	Worse
S7	79.58	80.46	Improved
S8	47.45	28.45	Worse
S9	27.48	27.56	Improved
S10	40.01	27.71	Worse
S11	38.22	21.5	Worse
S12	38.78	16.2	Worse
S13	79.84	65.23	Worse
S14	76.92	65.54	Worse
S15	74.36	64.79	Worse
S16	39.4	23.76	Worse
S17	36.25	14.9	Worse
S18	42.28	14.0	Worse

TABLE V. THE AVERAGE SAMPLE RUNS ON DWM, GDWM, AND MODER

Method	Precision	Recall	F1-Score	Balanced Accuracy
DWM	72.43	57.82	62.97	76.89
GDWM	84.78	68.62	71.47	84.31
ModER	80.308	44.24	51.691	72.113

TABLE VI. BENCHMARK RESULTS ON THE ABT-BUY DATASET

Model	F1-Score
DWM	10.83
Magellan	43.6
ModER	58.82

TABLE VII. BENCHMARK RESULTS ON THE ABT-BUY DATASET. BA IS THE BALANCED ACCURACY AND IS COMPUTED AS DESCRIBED IN TABLE I. THE INITIAL AND FINAL MODULARITY IS MEASURED ON THE FULL DISCONNECTED GRAPH AFTER ASSIGNING THE NEW MEMBERSHIPS AFTER THE COMPOSITE MODULARITY OPTIMIZATION STEP.

Beta	Sigma	Delta	# Nodes	# Single Nodes	# Edges	Initial Modularity	Final Modularity	Precision	Recall	F1-Score	BA
5	819	9	2173	274	3477	0.0827	0.4577	0.0182	0.3292	0.0345	0.6604
5	92	34	2173	854	2173	0.505	0.647	0.138	0.0367	0.058	0.5183
5	26	8	2173	273	3407	0.083	0.31	0.1669	0.1905	0.1779	0.595
2	92	8	2173	428	1611	0.2697	0.76134	0.6686	0.525	0.5882	0.7625

the current memberships and hypothetical random memberships, also known as the null model.

### C. The Interplay between Precision and Recall

Finally, we plot precision as a function of recall, also known as the precision-recall curve in Fig. 8. The difference is that

precision-recall curves are usually interpreted in the case of binary classification. In our entity resolution system, we are not assessing a binary classifier. We are, however, assessing cluster quality. Nevertheless, plotting precision as a function of recall on the 18 samples arranged in an ascending order shows that recall tends to be more stable than precision affirming our conclusion that ModER provides higher precision but not

Metrics across Synthetic Samples

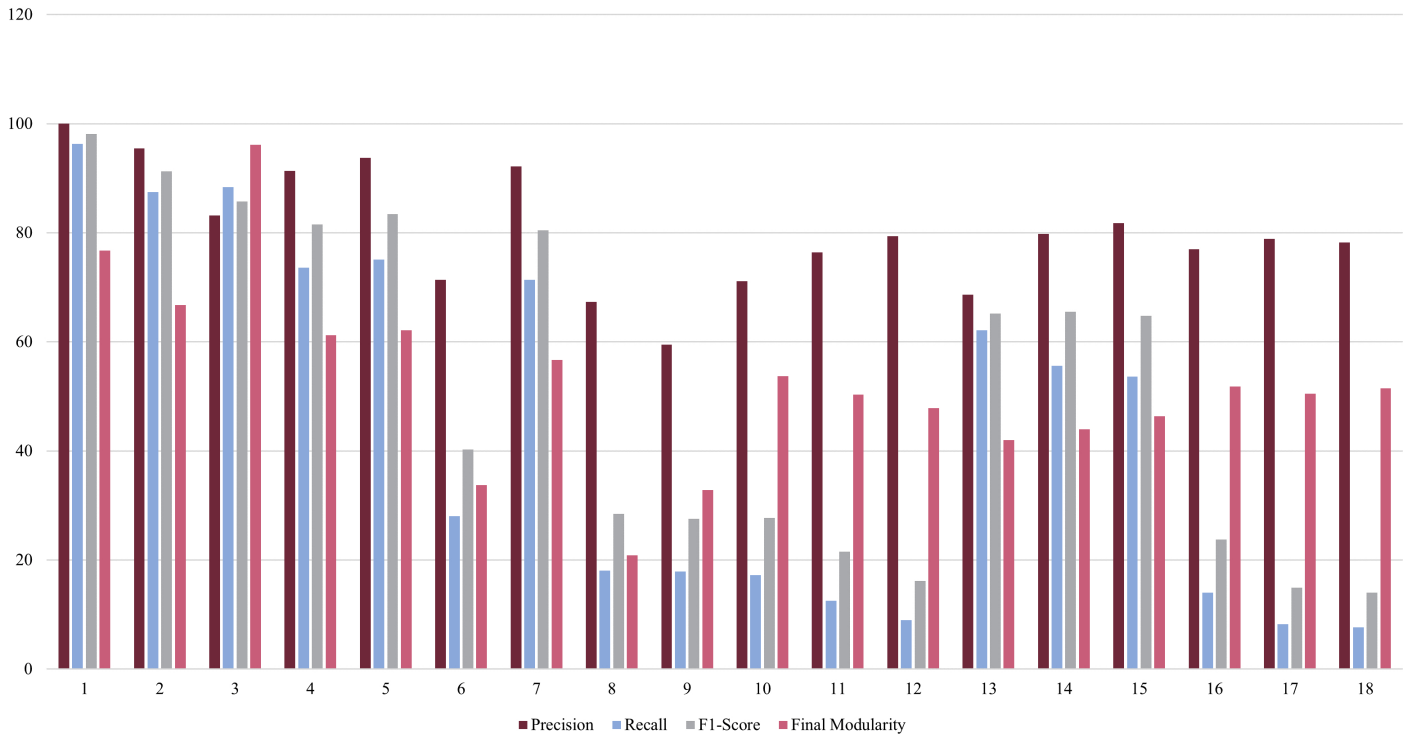


Fig. 6. A Bar Chart of Precision, Recall, f1-Scores, and Final Modularity after Running on ModER on the 18 Synthetic Samples.

necessarily recall.

#### D. Limitations

It is imperative and expected of course to understand the limitations of our approach here. First the 3 parameters that are controlled by the user  $\beta$ ,  $\sigma$  and  $\delta$  determine the sizes of the blocks. The size of the blocks eventually determines the sizes of the clusters in the modeled graph.

#### E. Takeaways and Future Work

In this work, we focused on the problem of unsupervised entity resolution for identifying unique entity profiles in ambiguous data. We defined ambiguous data as documents or records that do not adhere to a schema and come in unknown layouts and sometimes with inferior quality. In addressing this problem, We first focused on the meaning of similarity and matching. The problem lies in that if two documents are 100% similar, that does not mean that they refer to the same entity. In contrast, if 2 documents match at 10%, that does not mean that we are 100% sure that both do not refer to the same entity. Hence the inherent uncertainty and the nature of the problem. We started from the position that the only sure thing is that similarity between 2 documents of 50% is extreme uncertainty. Then assumed that below 50% of matches tend to ensure that the two documents do not refer to the same entity. On the other hand, two documents with a matching probability of more than 50% have some certainty that they might refer to the same entity. In the GDWM [15], we designed the system based on the later observation that higher mating probabilities should be detected with more certainty. While in

ModER, we generalized to all cases. That generalization came with some cost in performance, but it was not that significant, and compared with other methods in similar conditions, it gave respectable results. The point is that unsupervised entity resolution is a complex problem that needs to be addressed in a more sophisticated way. In addition, the concept of string similarity needs to be reconsidered as traditional similarity functions are the actual bottlenecks in this process. Some deep learning, machine learning, and graph approaches introduced learned similarities, such as in [13] and in [47]. In addition, this problem of unsupervised entity resolution could be generalized to other topics in natural language processing and information retrieval since it resembles the entity recognition problem and the search problem.

## VII. CONCLUSION

Here we introduced ModER, which stands for Modularity Composite Optimization for Entity Resolution, a framework combining multiple steps and algorithms. The method can quickly identify entity profiles in highly ambiguous data, overcoming the need to set matching thresholds. The method also limits user input to 3 parameters set using statistical percentile approximations. We based our work on the state-of-the-art unsupervised entity resolution, the DWM. In addition to the GDWM. To our knowledge, this technique has not been explored before. Our Composite Modularity Entity profiling step is innovative and can provide better results when benchmarked. In the future, we plan to address challenges such as the breakdown of high recall clusters even though they might not be imprecisely profiled. In addition, we address the memory-intensive bipartite approach posing a bottleneck for large

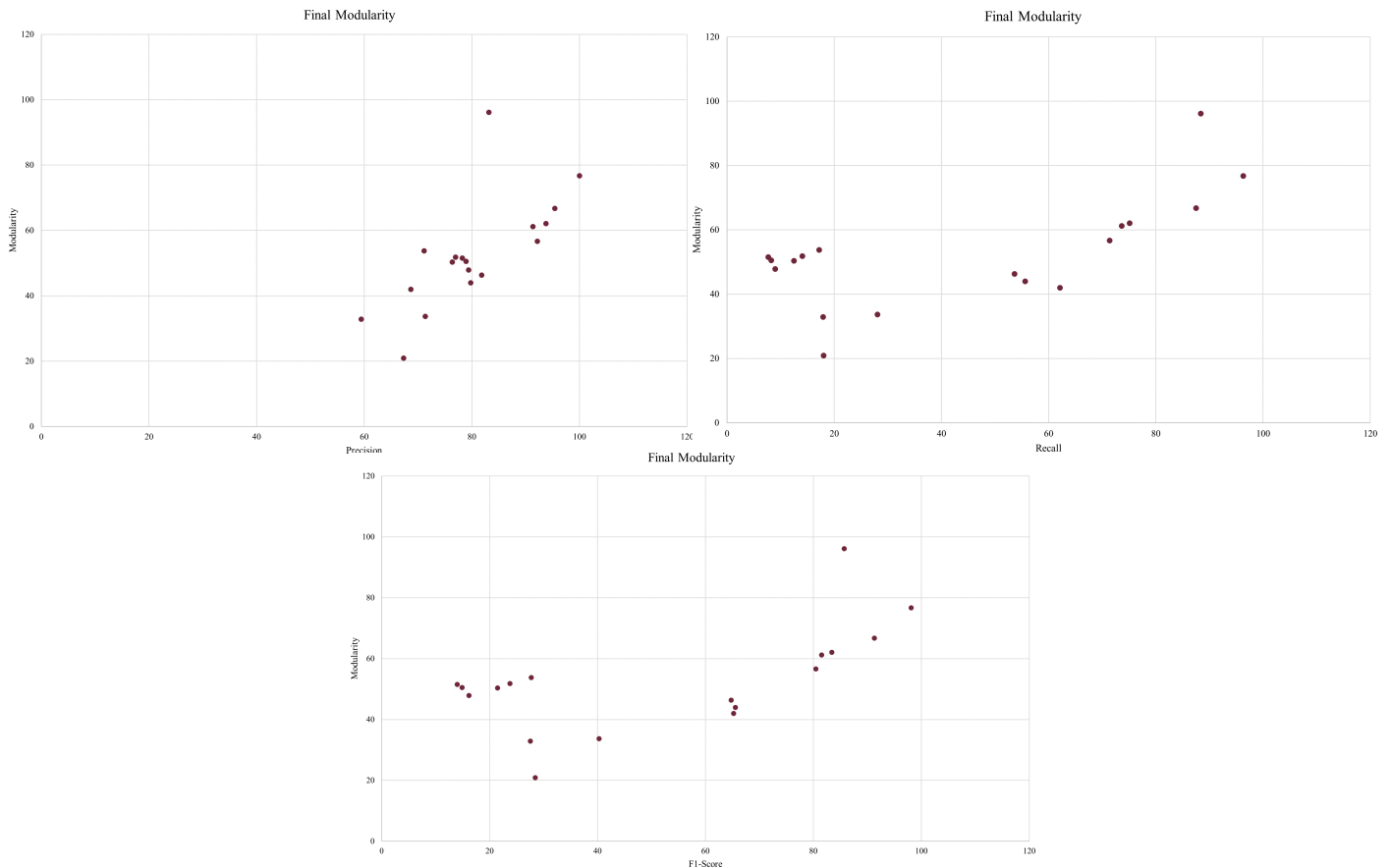


Fig. 7. A Scatter Plot of Values of Precision, Recall and f1-Score against Final Modularity Values on ModER after Running the 18 Synthetic Datasets.

profiles.

#### CONFLICT OF INTEREST STATEMENT

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

#### AUTHOR CONTRIBUTIONS

IAE created the ModER framework and wrote the article, JRT created the DWM framework and provided feedback and guidance on creating ModER as main PI, NH benchmarked the ModER framework and the GDWM framework and MAS tested the ModER framework and the GDWM framework on synthetic datasets and preprocessed the benchmark datasets.

#### FUNDING

This material is based upon work supported by the National Science Foundation under Award No. OIA-1946391.

#### REFERENCES

- [1] J. R. Talburt, D. Pullen, L. Claassens, R. Wang *et al.*, "An iterative, self-assessing entity resolution system: First steps toward a data washing machine," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 12, 2020.
- [2] D. G. Brizan and A. U. Tansel, "A. Survey of Entity Resolution and Record Linkage Methodologies," vol. 6, no. 3, p. 11, 2006.
- [3] J. R. Talburt and Y. Zhou, "A Practical Guide to Entity Resolution with OYSTER," in *Handbook of Data Quality: Research and Practice*, S. Sadiq, Ed. Berlin, Heidelberg: Springer, 2013, pp. 235–270. [Online]. Available: [https://doi.org/10.1007/978-3-642-36257-6\\_11](https://doi.org/10.1007/978-3-642-36257-6_11)
- [4] A. K. Elmagarmid, P. G. Ipeirotis, and V. S. Verykios, "Duplicate Record Detection: A Survey," *IEEE Transactions on Knowledge and Data Engineering*, vol. 19, no. 1, pp. 1–16, Jan. 2007, conference Name: IEEE Transactions on Knowledge and Data Engineering.
- [5] V. I. Levenshtein *et al.*, "Binary codes capable of correcting deletions, insertions, and reversals," in *Soviet physics doklady*, vol. 10, no. 8. Soviet Union, 1966, pp. 707–710.
- [6] M. S. Waterman, T. F. Smith, and W. A. Beyer, "Some biological sequence metrics," *Advances in Mathematics*, vol. 20, no. 3, pp. 367–387, 1976.
- [7] T. F. Smith, M. S. Waterman *et al.*, "Identification of common molecular subsequences," *Journal of molecular biology*, vol. 147, no. 1, pp. 195–197, 1981.
- [8] M. A. Jaro and V. C. Walker, *Unimatch: A record linkage system: Users manual*. The Bureau, 1978.
- [9] J. R. Ullmann, "A binary n-gram technique for automatic correction of substitution, deletion, insertion and reversal errors in words," *The Computer Journal*, vol. 20, no. 2, pp. 141–147, 1977.
- [10] A. E. Monge, C. Elkan *et al.*, "The field matching problem: algorithms and applications," in *Kdd*, vol. 2, 1996, pp. 267–270.
- [11] P. Jaccard, "Étude comparative de la distribution florale dans une portion des alpes et des jura," *Bull Soc Vaudoise Sci Nat*, vol. 37, pp. 547–579, 1901.
- [12] K. S. Jones, "A statistical interpretation of term specificity and its application in retrieval," *Journal of Documentation*, vol. 28, pp. 11–21, 1972.
- [13] D. Zhang, D. Li, L. Guo, and K. Tan, "Unsupervised Entity Resolution

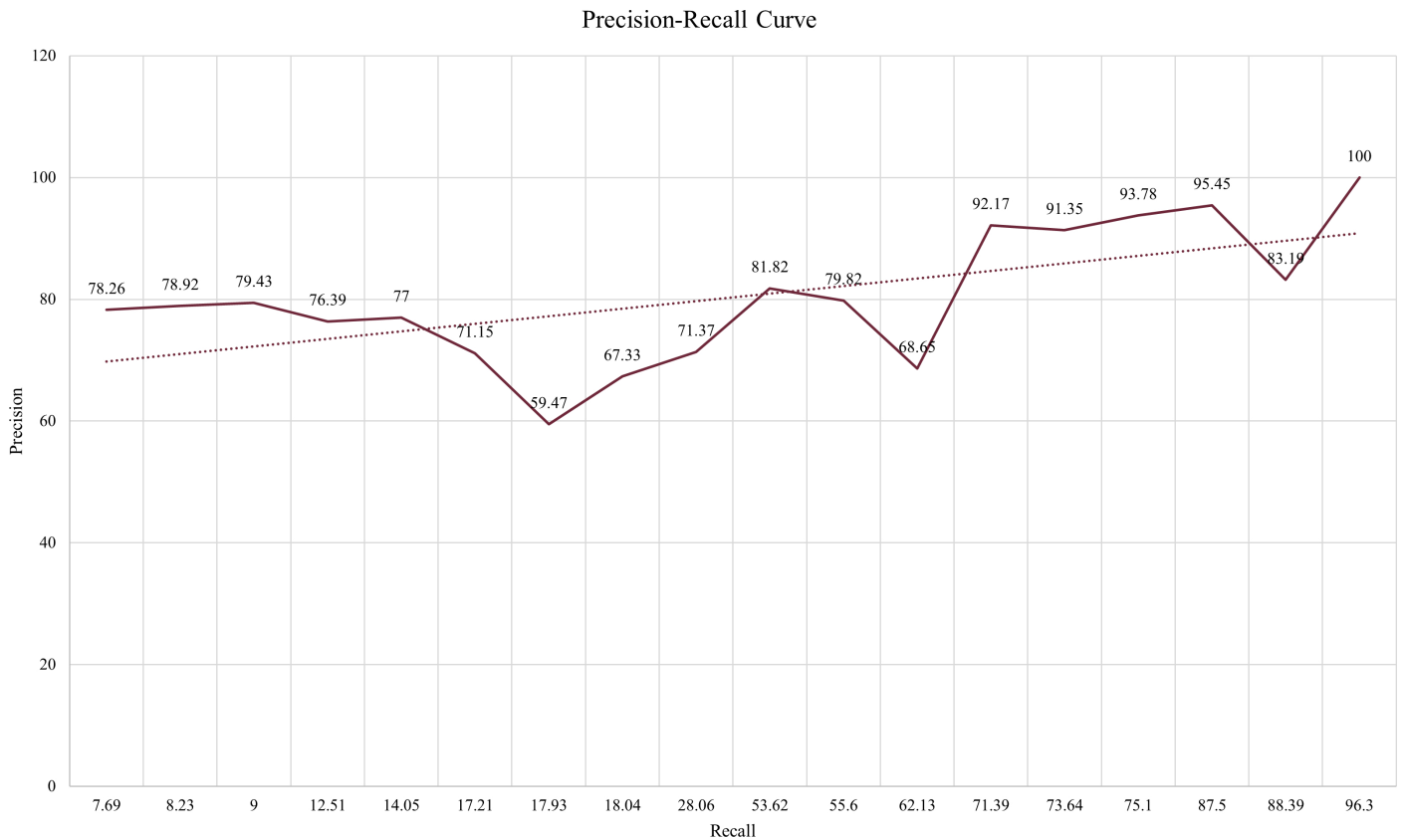


Fig. 8. Precision-Recall Curve.

with Blocking and Graph Algorithms,” *IEEE Transactions on Knowledge and Data Engineering*, pp. 1–1, 2020.

- [14] B. Li, W. Wang, Y. Sun, L. Zhang, M. A. Ali, and Y. Wang, “GraphER: Token-Centric Entity Resolution with Graph Convolutional Neural Networks,” *AAAI*, vol. 34, no. 05, pp. 8172–8179, Apr. 2020. [Online]. Available: <https://ojs.aaai.org/index.php/AAAI/article/view/6330>
- [15] I. A. Ebeid, J. R. Talburt, and M. A. S. Siddique, “Graph-based hierarchical record clustering for unsupervised entity resolution,” in *ITNG 2022 19th International Conference on Information Technology-New Generations*. Springer, 2022, pp. 107–118.
- [16] J. Wang, H. T. Shen, J. Song, and J. Ji, “Hashing for Similarity Search: A Survey,” *arXiv:1408.2927 [cs]*, Aug. 2014, arXiv: 1408.2927. [Online]. Available: <http://arxiv.org/abs/1408.2927>
- [17] A. Broder, “On the resemblance and containment of documents,” in *Proceedings. Compression and Complexity of SEQUENCES 1997 (Cat. No.97TB100171)*, Jun. 1997, pp. 21–29.
- [18] R. Wu, S. Chaba, S. Sawlani, X. Chu, and S. Thirumuruganathan, “Zeroer: Entity resolution using zero labeled examples,” in *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*, 2020, pp. 1149–1164.
- [19] Y. Li, J. Li, Y. Suhara, A. Doan, and W.-C. Tan, “Deep entity matching with pre-trained language models,” *arXiv preprint arXiv:2004.00584*, 2020.
- [20] O. Benjelloun, H. Garcia-Molina, D. Menestrina, Q. Su, S. E. Whang, and J. Widom, “Swoosh: a generic approach to entity resolution,” *The VLDB Journal*, vol. 18, no. 1, pp. 255–276, 2009.
- [21] G. Jeh and J. Widom, “SimRank: a measure of structural-context similarity,” in *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, ser. KDD ’02. New York, NY, USA: Association for Computing Machinery, Jul. 2002, pp. 538–543. [Online]. Available: <https://doi.org/10.1145/775047.775126>
- [22] F. Wang, H. Wang, J. Li, and H. Gao, “Graph-based reference table construction to facilitate entity matching,” *Journal of Systems and Software*, vol. 86, no. 6, pp. 1679–1688, Jun. 2013. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0164121213000484>
- [23] H. Wang, J. Li, and H. Gao, “Efficient entity resolution based on subgraph cohesion,” *Knowledge and Information Systems*, vol. 46, no. 2, pp. 285–314, 2016.
- [24] A. Saeedi, M. Nentwig, E. Peukert, and E. Rahm, “Scalable Matching and Clustering of Entities with FAMER,” *Complex Systems Informatics and Modeling Quarterly*, vol. 0, no. 16, pp. 61–83, Oct. 2018, number: 16. [Online]. Available: <https://csimq-journals.rtu.lv/article/view/csinq.2018-16.04>
- [25] U. Draisbach, P. Christen, and F. Naumann, “Transforming Pairwise Duplicates to Entity Clusters for High-quality Duplicate Detection,” *J. Data and Information Quality*, vol. 12, no. 1, pp. 3:1–3:30, Dec. 2019. [Online]. Available: <https://doi.org/10.1145/3352591>
- [26] L. Kolb, Z. Sehili, and E. Rahm, “Iterative computation of connected graph components with MapReduce,” *Datenbank-Spektrum*, vol. 14, no. 2, pp. 107–117, 2014, publisher: Springer.
- [27] N. Kang, J.-J. Kim, B.-W. On, and I. Lee, “A node resistance-based probability model for resolving duplicate named entities,” *Scientometrics*, vol. 124, no. 3, pp. 1721–1743, Sep. 2020. [Online]. Available: <https://doi.org/10.1007/s11192-020-03585-4>
- [28] L. Page, S. Brin, R. Motwani, and T. Winograd, “The PageRank citation ranking: Bringing order to the web.” Stanford InfoLab, Tech. Rep., 1999.
- [29] M. Sadiq, S. I. Ali, M. B. Amin, and S. Lee, “A Vertex Matcher for Entity Resolution on Graphs,” in *2020 14th International Conference on Ubiquitous Information Management and Communication (IMCOM)*, Jan. 2020, pp. 1–4.
- [30] D. Zhang, L. Guo, X. He, J. Shao, S. Wu, and H. T. Shen, “A Graph-Theoretic Fusion Framework for Unsupervised Entity Resolution,” in *2018 IEEE 34th International Conference on Data Engineering (ICDE)*, Apr. 2018, pp. 713–724, iSSN: 2375-026X.
- [31] M. E. J. Newman, “Modularity and community structure in networks,” *Proceedings of the National Academy of Sciences of the United States*

- of America, vol. 103, no. 23, pp. 8577–8582, Jun. 2006. [Online]. Available: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1482622/>
- [32] L. Hagen and A. Kahng, “New spectral methods for ratio cut partitioning and clustering,” *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 11, no. 9, pp. 1074–1085, Sep. 1992. [Online]. Available: <http://ieeexplore.ieee.org/document/159993/>
- [33] S. V. Ovchinnikov, “On the transitivity property,” *Fuzzy Sets and Systems*, vol. 20, no. 2, pp. 241–243, Oct. 1986. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/0165011486900801>
- [34] G. Navarro, “A guided tour to approximate string matching,” *ACM computing surveys (CSUR)*, vol. 33, no. 1, pp. 31–88, 2001.
- [35] A. Rajaraman and J. D. Ullman, *Mining of massive datasets*. Cambridge University Press, 2011.
- [36] E. Zhu, F. Nargesian, K. Q. Pu, and R. J. Miller, “LSH Ensemble: Internet-Scale Domain Search,” *arXiv:1603.07410 [cs]*, Jul. 2016, arXiv: 1603.07410. [Online]. Available: <http://arxiv.org/abs/1603.07410>
- [37] J. H. Burrows, “Secure Hash Standard,” DEPARTMENT OF COMMERCE WASHINGTON DC, Tech. Rep., Apr. 1995, section: Technical Reports. [Online]. Available: <https://apps.dtic.mil/sti/citations/ADA406543>
- [38] C. Y. Lee, “An algorithm for path connections and its applications,” *IRE transactions on electronic computers*, no. 3, pp. 346–365, 1961.
- [39] M. J. Barber, “Modularity and community detection in bipartite networks,” *Physical Review E*, vol. 76, no. 6, p. 066102, Dec. 2007, arXiv: 0707.1616. [Online]. Available: <http://arxiv.org/abs/0707.1616>
- [40] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, “Fast unfolding of communities in large networks,” *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2008, no. 10, p. P10008, Oct. 2008, arXiv: 0803.0476. [Online]. Available: <http://arxiv.org/abs/0803.0476>
- [41] H. Köpcke, A. Thor, and E. Rahm, “Evaluation of entity resolution approaches on real-world match problems,” *Proceedings of the VLDB Endowment*, vol. 3, no. 1-2, pp. 484–493, 2010.
- [42] J. R. Talburt, Y. Zhou, and S. Y. Shivaiah, “Sog: A synthetic occupancy generator to support entity resolution instruction and research.” *ICIQ*, vol. 9, pp. 91–105, 2009.
- [43] K.-N. Tran, D. Vatsalan, and P. Christen, “Geco: an online personal data generator and corruptor,” in *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, 2013, pp. 2473–2476.
- [44] Y. Ye and J. R. Talburt, “Generating synthetic data to support entity resolution education and research,” *Journal of Computing Sciences in Colleges*, vol. 34, no. 7, pp. 12–19, 2019.
- [45] P. Konda, S. Das, P. Suganthan G. C., A. Doan, A. Ardalan, J. R. Ballard, H. Li, F. Panahi, H. Zhang, J. Naughton, S. Prasad, G. Krishnan, R. Deep, and V. Raghavendra, “Magellan: toward building entity matching management systems,” *Proceedings of the VLDB Endowment*, vol. 9, no. 12, pp. 1197–1208, Aug. 2016. [Online]. Available: <https://dl.acm.org/doi/10.14778/2994509.2994535>
- [46] J. P. Mower, “Prep-mt: predictive rna editor for plant mitochondrial genes,” *BMC bioinformatics*, vol. 6, no. 1, pp. 1–15, 2005.
- [47] Y. Lin, H. Wang, J. Chen, T. Wang, Y. Liu, H. Ji, Y. Liu, and P. Natarajan, “Personalized Entity Resolution with Dynamic Heterogeneous Knowledge Graph Representations,” *arXiv:2104.02667 [cs]*, Apr. 2021, arXiv: 2104.02667. [Online]. Available: <http://arxiv.org/abs/2104.02667>

# Remote International Collaboration in Scientific Research Teams for Technology Development

## An Exploration of Team Culture and Efficiency

Sarah Janböcke<sup>1</sup>, Toshimi Ogawa<sup>2</sup>, Koki Kobayashi<sup>3</sup>

Ryan Browne<sup>4</sup>, Yasuki Taki<sup>5</sup>, Rainer Wieching<sup>6</sup>, Johanna Langendorf<sup>7</sup>

SARAH JANBÖCKE ODR GmbH on behalf of Tohoku University, 45219 Essen, Germany<sup>1</sup>  
Smart-Aging Research Center (S.A.R.C.), Tohoku University, Sendai 980-8575, Japan<sup>2, 3, 4, 5</sup>  
Business Informatics and New Media, University of Siegen, 57072 Siegen, Germany<sup>6, 7</sup>

**Abstract**—Scientific research teams often find themselves in remote working situations due to their internationality. Incredibly complex technological projects demand close collaboration and knowledge-sharing management. Remote teamwork, especially in cutting-edge scientific technology development, comes with various challenges that can negatively influence the overall team performance and commitment to the project. Within the EU-Japan (EU-/MIC-funded) project e-VITA, a consortium of 22 multidisciplinary partners and around 80 people work on research regarding a virtual assistant for healthy and active aging. We conducted qualitative data within the consortium after nine months of teamwork to understand the influence of collaboration on commitment, personal performance, efficiency, and work outcome. Based on this research's outcome, we built a framework for future scientific research projects and consortia to increase efficiency and quality of teamwork, thus researchers' well-being.

**Keywords**—Teamwork; technology development; international collaboration; scientific team performance; potential technology leverage; scientific commitment; team efficiency; team commitment; team performance; collaboration

### I. INTRODUCTION

Project e-VITA, e-VITA Virtual Coach for Smart Aging, is an EU-Japan project under the EU Horizon 2020 program and MIC funding regarding the Japanese Society 5.0 movement. 22 international partners research from 2021 to 2023 regarding new technologies and methods to help an aging society deal with specific problems of their older people. The aim is to combine sociological, medical, and technological excellence to produce an innovative coaching system based on the needs of older autonomous living adults. Thus, a virtual coaching system that can provide personalized recommendations and everyday help improve older adults' life quality in Europe and Japan while also delivering opportunities to SMEs and NGOs to derive knowledge, services, and products from this joint research force. The aimed impact scale is wide-ranged and ambitious for all partners and stakeholders in project e-VITA [1]. Thus, this project is a rare opportunity to research specific factors of scientific teamwork in complex technological research consortia, especially under the influence of the COVID pandemic and its specific influence on remote teamwork.

Team-wise we face a relatively rare challenge in project e-VITA. The team is brought together from different backgrounds without being orchestrated like an average team in, e.g., industry. There is no existing team that seeks an extension with a hiring process. The group is teamed together from various organizations and needs to get along no matter what; and it is faced with high expectations from the grant giver [2]. Building a team spirit in remote teams with no touching points is a rare situation [3]. A considerable challenge is establishing self-organizing sub-teams within the whole group [4]. The project e-VITA members come from a culture of waterfall hierarchy [5] and non-self-reliant work that needed to be changed to become a self-organizing team structure with agile aspects [4] to reach the complex aim of the project. When working in industry, we find a relatively clear understanding of the company, product, and job. In a research project like project e-VITA, the project start presents like a start-up without clear organizational structures [6] but also without a concrete product to gather around. It is a rather vague idea of what the research should look like compared to what a start-up business plan looks like when facing investors [7]. Installing rather formal business and strategy documents like an innovation roadmap [8], charter documents [9], and communicational guides were the first steps to meet the upcoming challenges in such a setting. The installment of a technological platform for data exchange, meeting organization, calendar set-up, and workstream organization in a remote setting [10] was also organized within the team and its members. Furthermore, the project e-VITA consortium coordinators tried to set up clear work structures [11] comparable to organizational structures in companies that were supposed to lead to more success without the expected friction losses in traditional and complex research projects [12] in science.

This research aims to collect data about the experienced work setting and culture significantly different from joint research projects. In the very first step, it is not the aim to quantitatively evaluate the used tools but to qualitatively get an impression on the work experience [13]. Positive work experiences are linked with employees' positive three-layered work commitment [14]. Furthermore, a high commitment is linked to more efficiency and qualitatively higher work outcomes [15]. Apart from wellbeing and health benefits due to a positive work and team culture [16], we aim to deeper



understand influencing factors and work on a recommendation framework for future research projects with similar complexity. Thus, our research approach is to understand which factors influence international collaboration and teamwork, if the experienced environment and climate influence the individual commitment and the quality of work outcomes, and which specific factors influence the personal performance.

## II. METHODOLOGICAL APPROACH

### A. Episodic Narrative Interview

We conducted 12 episodic narrative interviews with project e-VITA consortia members. The aim of an episodic narrative interview [17] is to better understand a phenomenon by generating individual stories of experience about that phenomenon. An episodic narrative interview participant provides nested narrative accounts of their experiences with a social phenomenon within the context of a bounded situation or episode. The episodic narrative interview is made to generate tightly focused, phenomenon-centered narratives reflective of bounded circumstances. We aimed to explore the deeper levels of experience linked to commitment and its effects and avoid the validity threat of social desirability by using a method that leads the interviewees to intuitive ways of reporting, in contrast, to merely answering explicit questions. We thoroughly followed the steps as presented by Alison Mueller.

### B. Interviews

We interviewed members that had to fulfill the delimitation criteria [17] of being part of the consortium for the whole period of nine months of bringing experiences from other research projects/ consortia, and of being actively involved in the project e-VITA in the specific episode in contrast to being a silent member that becomes active in later stages of the project. Furthermore, the sample was equally mixed from members of the EU and Japanese sides of the project. Thus, the primary interview language was English. To overcome possible language barriers, we also conducted five interviews in Japanese and professionally translated them to English for analysis purposes. The interviews were conducted remotely via Zoom without video streaming.

The 12 interviewees (Table I) were between 33 and 60 years old, with an average age of 46 years. Amongst the interviewees, we found six senior researchers from industry and science, four university professors, and two persons with high-ranking industry jobs (CEO/CTO). All interviewees have leadership experience ranging from 1 to 100 reports in technology science and industry, medical service and science, the housing industry, and political consultancy. Work experience ranged from 5 to 32 years at the interviews. All interviewees have a middle to high involvement in the researched project of project e-VITA.

### C. Analysis Process

As the method of Episodic Narrative Interview by Alison Mueller is relatively new and innovative, it does not offer extensive guidance regarding the used and proven analysis steps. We thus chose to be guided by the ideas of Grounded Theory Analysis and to follow the suggested three coding steps of open coding, axial coding, and selective coding to steadily

re-compare data and found phenomena to, in the end, derive a theoretical framework for the research questions of interest [18].

After the interviews were numerically coded to preserve anonymity, we mixed the Japanese and EU data by changing the numerical order to ensure an analyzing process without intercultural presumptions. We used five W-questions within the first coding step to define meaningful passages within the interviews and for the first theoretical abstraction. We focused on what was said, who was involved, what aspects were essential or influencing, why they were essential, and what solution was chosen for specific situations or problems. Thus, the aim was to detach the relevant passages from the overall interview to get an accurate impression of meaningful aspects not only in the context of one interview but in relation to the other interviews and relevant passages within.

Subsequently, we axially coded the defined text passages. We used the same codes to find connections, similarities, and differences. In an additional step of axial coding, we reduced our code system to capture different perspectives on particular issues. In a last coding step of selective coding, we started condensing our code system into a category network based on the found core categories from our previous coding steps.

We now were able to form theories and connections within a framework that could be the base for better cooperation in future international research projects.

TABLE I. INTERVIEWEE CHARACTERISTICS IN RANDOM ORDER FOR ANONYMITY, OWN DESIGN

Age	Position	Branch	Reporting Employees	Years of experience in expertise field
55	Senior Researcher	Real Estate Development	4	10
43	Senior Researcher	Research Institute	1-10	10
60	Professor	University	20-100	25
42	Senior Researcher	University	100	8
40	Assistant Professor	University	10	10
60	Professor	University	10-12	32
50	Manager	Start-Up	4	20
56	Senior Researcher	Research Institute	1	31
35	Senior Project Officer	Research Institute	1	5
39	Senior Researcher, Project Manager	Research Institute	2- 6	15
37	CTO	Start-Up/ Research Institute	4	14
33	Assistant Professor	University	7	4

#### D. Validity Threats / Methodological Limitations

The researchers of this study are part of the research object project e-VITA, i.e., potentially part of the phenomenon. This bears the danger of participant-answers according to the considerations of social desirability. We addressed this validity threat to meet the quality criteria by involving a supervision process to exclude the researchers own relevance system [19] from the data conduction and analysis phase during the research process, by using different interviewers, not only to meet language requirements and challenges but also to balance the personal factors that could arise social desirability answers. We focused on making the interviewees feel most comfortable to freely describe their experience with the phenomenon of interest. The validity threat of social desirability during the data conduction phase was also addressed by avoiding the video call and using a neutral screen whilst conducting the interview. Furthermore, we deliberately used the narrative interview style to lead the interviewees into phases of free talking and reminiscence without considering the interviewer and their relationship to each other [17] thus avoiding effects of social acceptance validity threats.

To distance ourselves from our own relevance system [19] during the analysis, we chose to present the data to a third-party researcher that was not involved in the project e-VITA so far, nor in planning the presented study or in conducting the data. The aim was to involve a perspective that adds an outer view on the data and results to avoid super exceeding expectations within the analysis [20] and research project of this study.

We considered a translator effect as another possible challenge [21] that we met by using the native language speaker on the JP side for data conduction. For the EU side we only used English as a common language for the interviews. We ruled out most of the common translator threats by using an algorithm-based translator software and a person fluent in both languages JP and English that professionally supported the study in the translation process.

### III. RESULTS

Within the following section we will present the found phenomena and directly compare them to the adjoining theoretical base. As no directly linked research can be found so far for our specific research questions in this application field, we draw links between adjoining fields and transfer them to our specific application interest. We combine the two steps of theoretical background/ comparison and result presentation for the sake of readability and length. We aimed our analysis to our above-mentioned research questions and could thus verify the following aspects as influencing factors for remote scientific work and international collaboration in research teams.

By the majority the interviewees addressed their need for change in various categories, but also their favor of certain aspects. Thus, we could define the topics communication, technical infrastructure/ remote work, organizational structure, personal information, cultural differences, commitment, workload, vision/ shared goal, personal development/ growth,

and shared values/ team cohesion as main categories for our analysis, i.e., most meaningful aspects for the interviewees.

The interviewees showed a great willingness to share deep insights of their experienced collaboration with us during the interviews. Throughout all interviews we could identify the most prominent topic, communication that was always addressed but was also always linked to all different categories mentioned above. Another specific phenomenon was the great wish to talk about commitment and to clarify specific forms of commitment throughout the whole consortium. The wish to enhance the organizational structure within huge projects like the researched one was also found in all interviews. Especially facing the affecting factors due to the COVID pandemic situation and remote working aspects left the interviewees with many expressed challenges.

#### A. Commitment / Shared Vision

Overall, respondents felt a strong commitment (compare [14], [15]) to the project and were motivated to achieve a good result. Especially the shared vision and shared common goal were named as important aspects to tackle the high complexity and workload of the project. However, many interviewees expressed the need to give more focus on a common vision and its communication within the whole team and to external partners and media streams.

“I think the positive thing is that we are very, very committed. So that's very new for me. And despite the fact we can't meet in the European countries or in Japan. ... Really, thanks to all the partners and namely the work coordinators, which are very, very involved and committed in this project, I think it is, this is also the guarantee of our success.” (Interviewee 6).

“I think that there is an overall goal in this project and that their people are working to, uhh, a lot of people working together on one goal with a certain amount of honesty and endurance and competence.” (Interviewee 2).

“...it would be good to find ways to at least in the beginning, to insist on this kind of vision.” (Interviewee 11).

The interviewees showed a solid normative and continuous commitment [14] when expressing the need to fulfill expectations to grant givers. Interestingly, they wished for more opportunities to expand their affective commitment [22] by getting deeper into the project's shared visions and getting deeply involved with their teammates throughout the whole consortium. Though all consortium members stemmed from different 'home' organizations and planned to build a research project like common in their field of work, they showed a strong interest in building an own organization for the project e-VITA. The tendency to form an own organization with all its effects like being committed to “one brand” gives a useful indication to the later framework but also to motivational aspects that can enhance innovative behavior for consortium members [7], [15], [23], [24].

#### B. Workload

The interviewees criticized the fact that they felt to only work for the deliverables of the project contract rather than the physical result, which aligns with the finding of a high

normative commitment mentioned above [25]. This sometimes put them under pressure and made them feel that the already high workload was even more significant or not feasible; they felt overwhelmed not capable of managing their own and other expectations. Thus, the interviewees expressed a decreased innovative capability aligned with first burnout tendencies [26]–[28].

In addition, some interviewees commented that they did not have the time to read through all the parts of the reports, even though they were interested in the progress of the other teams. The feeling of not being fully part of the team due to lacking information led the interviewees to want to enhance communication streams for deeper project involvement. This aligns with the wish to feel affectively committed [14], [22] to the project and gain a deeper understanding for the whole organization as well as the wish to be part of a 'bigger thing' to enhance self-efficacy [29], [30].

They also felt that the regular meetings of all the teams were too long and not very profitable because often everyone only gave their presentation, and there was barely any time for discussion and exchange. Thus, we can detect the need to deeper identify the leadership style and team cultural desires that are applied to the project. Discussion satisfaction among consortium members is a leading force in innovation behavior and employee satisfaction, furthermore a specific challenge in virtual teams [31].

The interviewees felt essential communication was missing; they expressed the need for an enhanced communication structure to cover specific information needs.

“...my first impression of this project for the initial months is that it has been very hard to do.” (Interviewee 4).

“(Person’s name) is struggling with a lot of deliverable workload.” (Interviewee 1).

“I’m sure it must be very difficult for the other researchers who are also working on top.

of their own jobs.” (Interviewee 3).

“I felt like I had to give 200% or 300% to finally I felt that ummm I was finally able to get an answer and wondered if I only resolved the issue because I had gone that far.” (Interviewee 1).

“I wish that days were 40, 48 hours long, but unfortunately they are not.” (Interviewee 5).

### C. Team Communication / Shared Goal

Some interviewees criticized that each team worked on its own and that there was too little collaboration as a whole group. As a result, a lot of knowledge was lost, even though the interviewees were basically interested in a team-wide overview.

“But the fact that we are working in silos we are working individually is not helping.” (Interviewee 9).

“But if we think that we need to cooperate, I found it really difficult to identify a cooperation with them.” (Interviewee 12).

“My impression is, that we haven't yet reached the point of real collaboration, which will become necessary in the future as we implement the system.” (Interviewee 7).

“Some do not work together at all. They don't know what. So European people do not know what the Japanese people do and the other way around...” (Interviewee 2).

In addition, each team within the consortium had different ways of working, so that it was difficult for non-team members to understand how the others worked and what insights could be gained.

“...it was always a bit of a feeling of not knowing what's going to happen next. All people on the same boat, so it was a bit like a kindergarten teacher to, uh, yeah, to, to take care of all the people involved in this project, are they all there? Are they going to be in the meeting? Are they doing their homework, so to speak?...” (Interviewee 2).

Again, we identify the consortium members' wish to be fully bound to the project, the desire to be given a broader base for their affective commitment. Obviously, this wish is connected to efficiency optimization but also to creating a work environment that offers wellbeing aspects [11], [22], [29], [32].

### D. Team Cohesion

The interviewees expressed a strong wish to feel as a whole and powerful team. They expressed the desire to be part of a big group capable of stemming this high workload and high complexity of such a technological research project. This is in alignment with the finding that the interviewees expressed the desire to find an environment that gives plenty of room for affective commitment in a work and research field that normally is rather conservative compared to brands that are classically connected to affective commitment like Adidas, Nike, Google, Apple etc. Interestingly the wish could be found without a cultural difference [33], [34].

“... a strong team is necessary, I think, to face the workload of project e-VITA.” (Interviewee 9).

“It's the team spirit. Yes, it's the team spirit and the team.” (Interviewee 6).

### E. Organizational Structure and Leadership

The interviewees wished for more apparent structures and task definitions, both within the team itself and across teams. They felt that time was lost, and the already significant amount of work was increased. More transparent structures would also enable more effective planning so that the workload and project flow could be managed continuously. This leads to the assumption to closely define work structures and streams in alignment with the team's socio-technical environment. This is true especially with regard to the used technological infrastructure that is applied to the project and the virtual team [35], [36].

“...a number of leaders and they all have a different way of a different style of leadership. Totally different. But there are a lot of people work with all of them or a couple of them, so they experience different ways of leadership, and some are more or less fair or more or less committed. Some are very committed.

So that's just a more heterogeneous way of leadership. And a lot of leaders next to each other.” (Interviewee 2).

“...if you're not involved as the leader. Because the project is so big and so many work packages are working next to each other simultaneously. You kind of get lost because you if you're not in every meeting that there is available, you will lose track of what's going on and other work packages. And therefore, you will not know what's going on there” (Interviewee 2).

“The needed structure to make all this thing run, um, well, of course, is necessary,…” (Interviewee 5).

“So, you have to do this strict organization, but also flexibility to change your objectives and your approach.” (Interviewee 10).

Some interviewees explained that they sometimes had too many tasks for which they lacked staff and skills. Hiring appropriate staff and giving room to these processes was expressed as a problem linked to the wish for enhanced organizational structures and better communication. Finding and keeping the right scientific personnel is well known as being a challenge. Especially the advanced skills needed in cutting edge technological projects demand a thorough selection process to ensure work quality results [37], [38]. Most managing persons in scientific projects were never trained in managing skills such as hiring and selection processes. These challenges were not mentioned but remained closely linked to this aspect. It is most likely that those persons demand a higher amount of time fulfilling the task of hiring than those who were specifically trained like industry managers and HR experts.

“Sometimes (team member's name) would ask me in meetings what I thought, but I couldn't say much, and I was a bit muddled, and I really didn't feel I could say much, even though I was the leader of Work Package...” (Interviewee 1).

“To be honest, there are many areas that are not my area of expertise,…” (Interviewee 4).

“In terms of my work, I have to deal with areas that I don't have the knowledge or experience to deal with,…” (Interviewee 8).

“...and we actually ask a temporary worker to do it for us.” (Interviewee 1).

“I think we need to ask a specialist for that, and we need to ask someone to support us in that area in the future.” (Interviewee 3).

We could also identify the wish to find a defined leadership style based on the findings. Interviewees described the current style and their wishes not only in terms of leadership in the project that defines certain tasks and work structures but also in motivation and guidance through the project's complexity. Thus, we can conclude the need to research appropriate leadership styles like transactional, transformational [39], or servant leadership [40] and follow basic principles based on the individual necessity of the team and project requirements (compare also [31]).

#### *F. Remote Work / Technical Infrastructure*

The ongoing COVID pandemic intensified many of the problems, as people could only work remotely with each other, which, on the other hand, is valid for many international projects even without the pandemic ongoing. This meant that the individual component was lost entirely for some interviewees because the international meetings could only take place online. They explained that as a result, they could not get to know their colleagues at all or only to a much lesser extent so that many aspects of communication such as facial expressions or the individual personality could not be conveyed. At the same time, a good team and support in the team were named several times as essential motivators to withstand these critical working conditions [7], [36], [41]. On the other hand, the remote work increased the wish to find enhanced organizational and communicational structures. The interviewees expressed several times the need to restructure classical work processes due to increased communication and alignment times to ensure efficient remote work [36], [42].

Furthermore, the interviewees described the wish for skilled personnel that deals with technical infrastructural questions that, on top of their research workload, needed to be tackled by themselves without the according expertise to do so. The remote working situation longed for specific technical solutions to ensure an efficient workflow, data exchange, video meetings, and a secure working environment for sensible information. According to the interviewees, the existing solutions on the market were not made for the specific context of scientific research projects, which left the members with many open issues that hindered their research work and work environment.

“I will say this is the first time that we are doing all the coordination in such a huge project, all by remote.” (Interviewee 5).

“...we didn't meet each other face to face also didn't help.” (Interviewee 9).

“I think it's something important. That the human interaction is it's important for the collaboration, for the cooperation and so on.” (Interviewee 9).

“...we never met face to face. And we could see that the...umm. It took more time, let's say to, to adjust our... there's an expression in xxx (my language) saying that we..., meaning that we have to adjust to the other person.” (Interviewee 9).

“...discovering that we can have a productive and efficient collaboration purely online was also quite a good thing.” (Interviewee 11).

“I also appreciate it a lot the selflessness in the Japanese partners in helping us...” (Interviewee 5).

“I really see a mutual help in this collaboration.” (Interviewee 10).

#### *G. Cultural Differences / Personal Information and Development*

The cultural differences between the teams were on the one hand seen as enrichment, but at the same time also led to some misunderstandings and problems, especially between the

European and the Japanese team. Since communication differs in many ways in all involved cultures, everyone had to adjust to each other first, which cost a lot of time that was not given within the planning structure. However, these misunderstandings could partly be solved through communication by individual team members explaining their behavior afterwards, thus creating an understanding. Again, this phenomenon is strongly linked to the expressed wish to rethink organizational structures and different communication behavior [36], [42].

“I think throughout the first nine months, it also changed a bit from the Japanese side. So, in the end, in the beginning, they were less vocal about their needs and also about their limitations. Now they are more vocal about it. So, they sometimes say they, I'm just referring to specific institutions I had contact with. ... I'm sorry, I don't know anything about it. Someone else has to do it. ... From my side point of view, they feel more comfortable to tell us if something is just not possible for them with regards to schedule, competence, or anything like that.” (Interviewee 2).

“... this kind of approach is something that has never been seen before in Japan, especially in technical projects...” (Interviewee 8).

“Challenges also is that we come from a different cultural context. ... But I feel like it's a cultural thing in Japan. Maybe I am not sure. Maybe they need more time to get the approval of maybe the hierarchy.” (Interviewee 9).

“...first difficulty was to try to understand each other, especially with our Japanese counterpart, because there were some small difficulties at the beginning and understanding each other.” (Interviewee 10).

“...it's not the same with the Japanese partner. Sometimes I think that the communication channels, it's completely different between us and them...” (Interviewee 12).

“There are three or four different European cultures that are packed together and in one side of the project, I would say, and it felt like the Japanese were between themselves, more in line with what they were doing then than the European side.” (Interviewee 11).

“That experience itself is something that I had never experienced before in my involvement with domestic projects in Japan. ... it was the first time that I had actually experienced this kind of emphasis in a project, and I think it was a great experience for me to be exposed to the values of this kind of team.” (Interviewee 8).

“I also started to understand how people work in such a mixed project and their habits.” (Interviewee 7).

As mentioned above, it was often expressed that more effective communication was desired. Particularly through remote work, some of the interviewees felt that communication was essential to create a team feeling, get to know each other and work effectively together. According to some of the interviewees, work-related conversations should occur more often and be shorter. On the other hand, the personal component should be strengthened by creating a framework for conversations without a work-related context. This aligns with

the idea to find an environment that leaves more room for affective commitment and its effects [43], [44].

“In the end, the only way to get along with a group of people who don't know each other is to talk to them. That's all there is to it.” (Interviewee 7).

“...it was really important to have these series of meetings and conferences. ... I think that this should be an added value for the future and which we will have more time and more space on board to talk together and to plan together. ... we really need to be in communication with more partners and also from the Japanese counterpart. That is something that we can do” (Interviewee 10).

“...because for myself, I think it's you can deal with everything if you talk about it. So, if somebody has someone has a hard time in his private life or I don't know and he can't do his parts at us, that's not a problem itself, because then just tell me and we can work around it.” (Interviewee 2).

Based on the various and extensive aspects that we gained throughout the interviews, we could identify two main pillars with a respective substructure that will be discussed in the following framework derivation section.

#### IV. FRAMEWORK DERIVATION

We derived a theoretical framework from the coded data and found the according information within the last step of theoretical abstraction. The framework will be a mixture of textual and visual overviews that shall aim to build recommendations to further contexts of remote scientific work.

We could identify two main pillars that headline the discussed categories – terms of commitment and organizational structure. Within standard research project organization, we find a discussion focus in research teams and their organizations on content-based aspects like research topic, research question, grant giver restrictions, deadlines, and deliverables (Fig. 1). Discussions around structural aspects and terms of commitment seem to be missing, which initiated the expressed need for change from the interviewees.

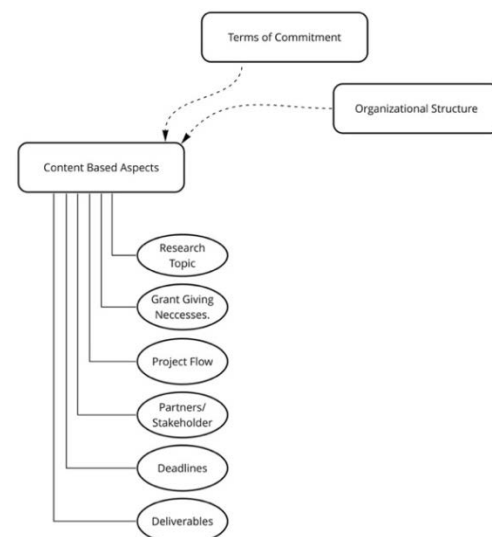


Fig. 1. Commonly Discussed Project Aspects and Missing Factors, Own Design.

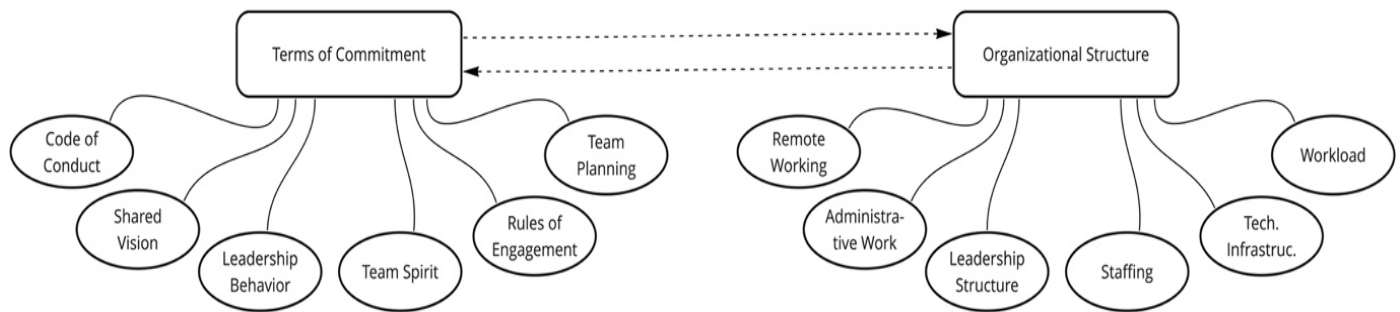


Fig. 2. Framework Extension Visualization, Own Design.

Triggered by the interview style, we could identify meaningful aspects that the interviewees either wished to be realized or intensified. We thus could build Fig. 2 from which we can derive questions that future research teams in similar working situations like project e-VITA should address before starting the project (Fig. 2).

Under Terms of Commitment, we could identify six subcategories addressed with the interviewees' wish of change.

- 1) Code of Conduct
- 2) Shared Vision
- 3) Leadership Behavior/ Responsibilities
- 4) Team Spirit
- 5) Communicational Engagement/ Rules of Engagement
- 6) Team Planning

For Organizational Structure, we could identify six further subcategories.

- 7) Remote Working Aspects
- 8) Administrative Work and Timing
- 9) Leadership Structure
- 10) Staffing Issues
- 11) Technological Infrastructure and Maintenance
- 12) Workload

We will now present the questions derived from the interview material, which should be addressed in advance before starting a research project within the whole consortium. Based on the defined categories, the interviewees described meaningful aspects and questions, topics, and problems they favor to be addressed. We summarize those as follows.

1) Code of conduct / Shared values

a) How do we want to deal with competition? Knowledge-sharing?

b) To which extent do we integrate cultural differences as a beneficiary factor? Or do we just ignore them? Will we use them as learning aspect for personal development? Do we explicitly address them before or while working?

c) How do we deal with motivational aspects? Who is responsible for motivating the team, each one on his/her own or the consortium leader? How do we deal with missing motivation?

2) Shared vision

a) Which goals can we identify? Is it one goal for all, or can we combine various goals? Are there individual goals that are hard to integrate into the whole project? What do we want to achieve after our joint project time? Which result do we want to see in the end?

b) Is there any sort of a "brand message" that we can describe for our project, may be based on its name? For example: xxx (name of the project) stands for...

3) Leadership behavior/ Responsibilities

a) Do we have a leadership structure that can be clearly defined? Which responsibilities do we see for our possible leadership? Can we share responsibilities and leadership workload?

b) Which leadership style do we want to apply?

c) Who is filling the roles that we defined for our leadership?

4) Team spirit

a) How do we plan to work together? Do we see the project as a joint project by people from various organizations that meet every now and then? Do we want to build our own very close team, like our own little organization?

b) Can we all commit to the shared team spirit goals? How do we deal if individual positions do not align with our overall team spirit goal?

5) Communicational engagement

a) Can we define communicational rules? Which will be essential to us?

b) Which communication channels do we want to use? Chat, Mail, Telephone etc.

c) Can we define reaction times to different communicational media streams like WhatsApp, Mail, Chat, Kanban boards, phone calls, ToDo lists etc.?

d) Can we rate the consequences of chosen communication media streams? Can we use them and commit to our communicational rules and reaction times?

e) Can we define a timing range and content that needs to be discussed regularly? How often? When?

6) Team planning

a) Do we want to be one team or act in specific silos? Do we want to experience the benefits of getting to know researchers with diverse backgrounds, and how do we manage

this? How can we benefit from diverse backgrounds and integrate them?

b) How much time do we plan to set up and proceed with team planning sessions?

c) How do we deal with positions that do not align with our overall goal regarding team planning sessions?

7) *Remote working aspects*

a) Which experiences do we have with working exclusively remote? Which aspects of teamwork are essential and need to be included in the project?

b) Which benefits can we identify, which challenges? Which differences in contrast to face-to-face work do we need to consider?

c) Do we plan additional time to tackle the identified challenges?

d) How do we deal with missing answers to possible upcoming challenges?

e) How do we deal with knowledge/ result sharing aspects throughout the internal teams? How can we tackle information overflow vs. missing information? How do we share results, papers other information in a manageable way?

8) *Administrative work and timing*

a) How much administrative work do we expect from this project?

b) Which experience from previous projects can we share? Are they beneficial to our situation now?

c) Can we plan additional time for administrative work?

d) Who will be responsible for tackling administrative issues? Who is in charge, and can we delegate tasks?

9) *Leadership structure*

a) Can we define the leadership structure that we previously discussed in question 3a?

b) Can we sketch the leadership structure in one organigram that will be mandatory for all members?

c) How do we deal with changes in our project structure and according responsibilities?

d) How do we deal with missing commitment?

10) *Staffing issues*

a) Do we have the needed competencies already onboard, or do we need to expand?

b) Did we plan enough time to find the fitting team extensions and competencies?

c) How do we deal with missing competencies? How do we close possible gaps?

11) *Technological infrastructure and maintenance*

a) Which technical infrastructure do we want to use to work remotely? Which technological solution/ platform for which task discussed in 5b+c?

b) Which experiences can we share from other projects?

c) Who is responsible for setting up the technical infrastructure?

d) Who will maintain the chosen solution throughout the whole project, and do we have, or these persons have enough resources for fulfilling their task?

12) *Workload*

a) Did we realistically estimate the upcoming workload?

b) Can we identify gaps, challenges, overloads? How do we handle them?

c) Did we realistically plan the necessary time to tackle our workload? If not, how do we deal with upcoming problems?

d) How do we deal with various positions about workload manageability? Especially about aspects discussed in 1c?

Based on our findings we argue that beside content-based aspects future scientific research projects especially in remote working situations should address Terms of Commitment and Organizational Structure aspects to ensure efficiency, optimal team performance and researchers' wellbeing and commitment to the project. All three will be decisive for the later project's result and should stand equally beside each other in terms of importance (Fig. 3). This equality can be derived from our interview data. Thus, our orientation framework is of interest for future teams and grant giving stakeholders that review and proof project proposals, project flows, and results.

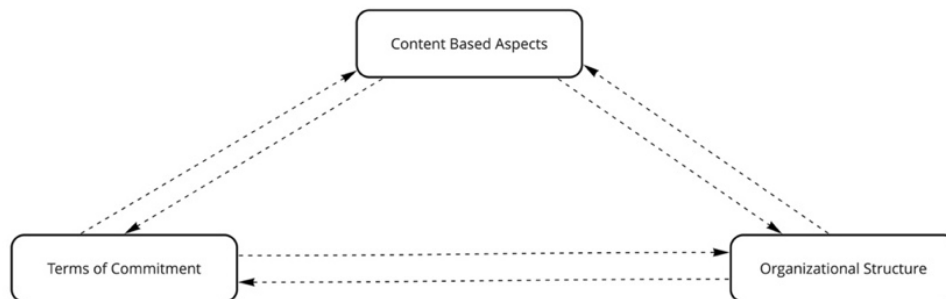


Fig. 3. General Framework Overview, Own Design.



## V. DISCUSSION

The presented study shows the need for additional planning when orchestrating a research project, especially in remote scientific, international collaboration. The high complexity of cutting-edge technological research projects is an additional complicating factor for the work environment. Based on the conducted data, we could show that the need to introduce additional aspects apart from content-based discussions is strongly given. Doing so will most likely positively influence team commitment and thus personal and team performance, efficiency, and work quality.

### A. Affective Commitment

International remote collaboration seems to be determined by many more soft factors than considered when planning projects. However, the team and individual performance can be positively influenced by the conscious planning of the soft framework factors. The possibility of living affective commitment in the projects is closely related to increased efficiency [45], [46]. In contrast to normative and continuous commitment, affective commitment is known to cause performance to skyrocket. A factor that should not be neglected in difficult working conditions such as remote work. A variety of framework conditions give the possibility of being able to form affective commitment. The interviewees described which organizational-psychological measures can be taken to bond emotionally to the project and thus create an environment of well-being. The sub-items of the found Terms of Commitment and the items of Organizational Structure are therefore sub-areas of the overall measure to create an environment for affective commitment. Employee commitment is widely researched for companies and their working environments.

In connection with employee identification, employee retention, and performance improvement, commitment is essential for regulating personnel processes. Organizational commitment, in general, is known as a critical driver to motivation and performance improvement of employees. However, it is advisable to distinguish the specific forms of commitment to deeply understand the influencing factors [46]. Continuous commitment is based on employees' cost-benefit calculation, i.e., leaving the company is associated with economic disadvantages for an individual that exceed the leaving benefit, a bond is created, and the employee tends to remain in the company. In contrast, normative commitment is based on an individual's values and perceived moral obligation to remain loyal to his or her company. The employee might feel committed to the company due to favors his/her superior might have given him/her in the past. On the other hand, affective commitment is defined by Meyer and Allen as a solid psychological bond that ties the individual to the organization. It describes the emotional attachment to the organization and has its origin in positive experiences with the company [47]. Affective commitment is declared to be the most potent form of employee loyalty to the organization and is not moderated by cultural country specifics [48]. A high level of affective commitment is associated with increased motivation, the will to take additional tasks, and employees' feelings of joy and pride for their organizational affiliation. Thus, affective

commitment correlates positively with higher individual performance and efficiency [49].

The found sub-items of our two main pillars, Terms of Commitment and Organizational Structure, describe with high agreement what organizational commitment research suggests applying for employee commitment improvement. Without explicitly knowing and naming aspects of what organizational research has known for decades, the interviewees described them for an entirely different setting. We argue that transferring organizational knowledge in the form of our framework to scientific research projects will increase affective commitment for remote researchers and the academic world and thus performance, efficiency, and researchers' well-being.

### B. Limitations

As discussed in the methodical section, we thoroughly tried to rule out possible limitations to the methodological approach. However, we cannot exclude the possibility that we did not get the full emotional range from the interviewees based on their cultural background or other personal influencing factors. This might give room for the fact that we did not record specific problems to the same extent as those mentioned by the persons who talked utmost limitation-free. However, this is true for every qualitative study since the researcher can never be entirely sure if the interviewees provide their whole knowledge or feelings. We thus trust our interviewees' expertise and professionalism, thus their given insight.

We mixed insights from university employees with those working in the industry. We deliberately did not separate those two sectors to gain a broad overview of multidisciplinary project structures in which industry and science mix. Knowing that the used work environment and standards most likely differ from science to industry, we consider a mixed sample approach as the most realistic one when seeking a framework for projects that combine science and industry partners most of the time.

Since we conducted data in Japan and Europe, specifically Germany, France, Italy, and Belgium, we cannot predict our conclusions transferability to other multi-cultural settings. However, some of the interviewees have extensive international experience, so we assume that the resulting framework can be adapted to other cultural standards and demands. We formulated the framework questions in an open manner that leaves enough room for individual cultural adaptations. Furthermore, since the suggested framework focusses mainly on increasing employees' affective commitment, we can also minimize the risk of cultural moderating factors [48].

We interviewed mainly persons with higher responsibility, i.e., higher hierarchical position. We cannot entirely be sure if the meaningful aspects found can be transferred to the emotional narratives of low-ranking employees. We considered the interviewed persons as experts for their field, including their lower-ranking staff. However, we cannot entirely rule out the difference in findings when replicating the study with a different sample and adapted delimitation criteria.

Deliberately we excluded researching the technological tools used and mentioned to ensure the technological infrastructure for remote work. This study focused on meaningful aspects to participants of remote international research projects, not on evaluating the technological solutions used.

### C. Future Studies

For the future, we see the open question of how the resulting framework described above can be implemented and used for prospective research projects. We see the danger that apart from focusing on the necessary content-based aspects, time is limited to concentrate on Terms of Commitment and Organizational Structure. Especially at the beginning of a project, when a consortium starts to find a joint base, additional time for such workshops might be missing or not considered relevant. After nine months of the project, the described challenges, obstacles, and problems are prominent for the interviewees. They might not have been prominent in the very beginning. However, since most partners in such projects come with experience from other research consortia, we can assume that the mere trigger to spend time on additional planning aside from the content-based aspects gives the suitable indication and priority. It might seem to add additional work at the beginning of a project. The possible reactance towards that must be overcome to later benefit from the positive effects of such a framework application. Especially the consortium leaders will oversee transporting the necessity and creating room for workshops, discussions, and fixation of framework questions like those mentioned above.

Next, we see a necessary evaluation of the framework. As a first step, we regard the opportunity to discuss the found framework and questions with experienced researchers and members from various projects and possibly adapt our recommendation according to those findings. A comparing evaluation might be possible for the far future, i.e., comparing projects and their results with and without applied framework.

### ACKNOWLEDGMENT

This work was partially supported by the joint EU and MIC project e-VITA. Project e-VITA received funding from the European Union H2020 Program under grant agreement no. 101016453. The Japanese consortium received funding from the Japanese Ministry of Internal Affairs and Communication (MIC), Grant no. JPJ000595. Special acknowledgment to the members of the project e-VITA consortium and Tohoku University, Smart-Aging Research Center, for their support.

### REFERENCES

- [1] Homepage of e-VITA - e-VITA Virtual coach. (2021). <https://www.e-vita.coach>.
- [2] H. Liu, S. Gao, H. Xing, L. Xu, Y. Wang, and Q. Yu, "Shared leadership and innovative behavior in scientific research teams: a dual psychological perspective," *Chinese Management Studies*, Jan. 2021, doi: 10.1108/CMS-02-2020-0070.
- [3] A. Malhotra, A. Majchrzak, and B. Rosen, "Leading virtual teams," *Acad. Manag. Perspect.*, vol. 21, no. 1, pp. 60–70, Feb. 2007, doi: 10.5465/AMP.2007.24286164.
- [4] N. B. Moe, T. Dingsøy, and T. Dybå, "Understanding self-organizing teams in agile software development," *Proc. Aust. Softw. Eng. Conf. ASWEC*, pp. 76–85, 2008, doi: 10.1109/ASWEC.2008.4483195.
- [5] S. Ashmore, A. Townsend, S. DeMarie, and B. Mennecke, "An exploratory examination of modes of interaction and work in waterfall and agile teams," *Int. J. Agil. Syst. Manag.*, vol. 11, no. 1, pp. 67–102, 2018, doi: 10.1504/IJASM.2018.091361.
- [6] J. C. Picken, "From startup to scalable enterprise: Laying the foundation," *Bus. Horiz.*, vol. 60, no. 5, pp. 587–595, Sep. 2017, doi: 10.1016/J.BUSHOR.2017.05.002.
- [7] T. Nordgreen et al., "Challenges and possible solutions in cross-disciplinary and cross-sectorial research teams within the domain of mental health," *J. Enabling Technol.*, vol. 15, no. 4, pp. 241–251, Oct. 2021, doi: 10.1108/JET-03-2021-0013/FULL/PDF.
- [8] W. Eversheim, F. Brandenburg, T. Breuer, M. Hilgers, and C. Rosier, "Die InnovationRoadMap-Methodik," *Innov. für Tech. Produkte*, pp. 27–131, 2003, doi: 10.1007/978-3-642-55768-2\_3.
- [9] M. Crawford, "Defining the Charter for Product Innovation," *Sloan Manage. Rev.*, vol. 22, no. 1, 1980, Accessed: Feb. 07, 2022. [Online]. Available: <https://www.proquest.com/openview/2c03cd7b96dc4ba03652c5b24066038c/1?pq-origsite=gscholar&cbl=35193>.
- [10] E. Brynjolfsson, D. Rock, J. Horton, A. Ozimek, G. Sharma, and H. Y. T. Ye, "COVID-19 and Remote Work: An Early Look at US Data," *Natl. Bur. Econ. Res.*, pp. 1–16, 2020, Accessed: Jul. 31, 2020. [Online]. Available: [https://john-joseph-horton.com/papers/remote\\_work.pdf](https://john-joseph-horton.com/papers/remote_work.pdf).
- [11] S. Kauffeld, *Arbeits-, Organisations- und Personalpsychologie für Bachelor*. 2. Auflage. 2014.
- [12] B. R. Staats, K. L. Milkman, and C. R. Fox, "The team scaling fallacy: Underestimating the declining efficiency of larger teams," *Organ. Behav. Hum. Decis. Process.* vol. 118, no. 2, pp. 132–142, Jul. 2012, doi: 10.1016/J.OBHDP.2012.03.002.
- [13] P. E. Tesluk and R. R. Jacobs, "Toward an Integrated Model of Work Experience," *Pers. Psychol.*, vol. 51, no. 2, pp. 321–355, Jun. 1998, doi: 10.1111/J.1744-6570.1998.TB00728.X.
- [14] N. J. Allen and J. P. Meyer, "The measurement and antecedents of affective, continuance and normative commitment to the organization," *J. Occup. Psychol.*, vol. 63, no. 1, pp. 1–18, Mar. 1990, doi: 10.1111/j.2044-8325.1990.tb00506.x.
- [15] Z. Md, M. Sambasivan, and J. Johari, "The influence of corporate culture and organisational commitment on performance," *J. Manag. Dev.*, vol. 22, no. 7–8, pp. 708–728, Oct. 2003, doi: 10.1108/02621710310487873.
- [16] N. R. Baptiste, "Tightening the link between employee wellbeing at work and performance: A new dimension for HRM," *Manag. Decis.*, vol. 46, no. 2, pp. 284–309, 2008, doi: 10.1108/00251740810854168.
- [17] R. A. Mueller, "Episodic Narrative Interview: Capturing Stories of Experience With a Methods Fusion," *Int. J. Qual. Methods*, vol. 18, pp. 1–11, 2019, doi: 10.1177/1609406919866044.
- [18] J. M. Corbin and A. Strauss, "Grounded theory research: Procedures, canons, and evaluative criteria," *Qual. Sociol.*, vol. 13, no. 1, pp. 3–21, 1990, doi: 10.1007/BF00988593.
- [19] J. Kruse, "Qualitative Interviewforschung: ein integrativer Ansatz," *Grundlagentexte Methoden*, 2015.
- [20] B. Gransche, *Vorausschauendes Denken - Philosophie und Zukunftsforschung jenseits von Statistik und Kalkül*. Bielfeld: transcript, 2015.
- [21] I. Kapborg and C. Berterö, "Using an interpreter in qualitative interviews: does it threaten validity?," *Nurs. Inq.*, vol. 9, no. 1, pp. 52–56, 2002, doi: 10.1046/J.1440-1800.2002.00127.X.
- [22] L. Rhoades, R. Eisenberger, and S. Armeli, "Affective commitment to the organization: The contribution of perceived organizational support," *J. Appl. Psychol.*, vol. 86, no. 5, pp. 825–836, 2001, doi: 10.1037/0021-9010.86.5.825.
- [23] G. J. Tellis, J. C. Prabhu, and R. K. Chandy, "Radical Innovation across Nations: The Preeminence of Corporate Culture," <https://doi.org/10.1509/jmkg.73.1.003>, vol. 73, no. 1, pp. 3–23, Jan. 2009, doi: 10.1509/JMKG.73.1.003.
- [24] E. Traut-Mattausch, D. Frey, T. Greitemeyer, and B. Streicher, "Psychologie der Innovationen in Organisationen," 2006.
- [25] J. P. Meyer and N. M. Parfyonova, "Normative commitment in the workplace: A theoretical analysis and re-conceptualization," *Hum.*

- Resour. Manag. Rev., vol. 20, no. 4, pp. 283–294, Dec. 2010, doi: 10.1016/J.HRMR.2009.09.001.
- [26] A. Antonovsky, *Unraveling the mystery of health: How people manage stress and stay well*. Jossey-Bass, 1987.
- [27] A. J. Dubinsky, R. E. Michaels, M. Kotabe, C. U. Lim, and H. Moon, “Influence of Role Stress on Industrial Salespeople’s Work Outcomes in the United States, Japan and Korea,” *J. Int. Bus. Stud.* 1992 231, vol. 23, no. 1, pp. 77–99, Mar. 1992, doi: 10.1057/PALGRAVE.JIBS.8490260.
- [28] K. I. Ohbuchi, M. Suzuki, and Y. Hayashi, “Conflict management and organizational attitudes among Japanese: individual and group goals and justice,” *Asian J. Soc. Psychol.*, vol. 4, no. 2, pp. 93–101, Aug. 2001, doi: 10.1111/J.1467-839X.2001.00078.X.
- [29] M. Seligman and E. Diener, “Beyond Money- Toward an Economy of Well-Being,” *Psychol. Sci. public Interes.*, vol. 5, no. 1, pp. 1–31, 2004, [Online]. Available: <http://www.jstor.org/stable/40062297>.
- [30] Y. Ueda, “Organizational citizenship behavior in a Japanese organization: The effects of job involvement, organizational commitment, and collectivism,” *J. Behav. Stud. Bus.*, vol. 4, no. 1, 2011.
- [31] R. Huang, S. Kahai, and R. Jestic, “The contingent effects of leadership on team collaboration in virtual teams,” *Comput. Human Behav.*, vol. 26, no. 5, pp. 1098–1110, Sep. 2010, doi: 10.1016/J.CHB.2010.03.014.
- [32] R. J. Burke, S. Moodie, S. L. Dolan, and L. Fiksenbaum, “Job Demands, Social Support, Work Satisfaction and Psychological Well-Being among Nurses in Spain,” *SSRN Electron. J.*, Jul. 2012, doi: 10.2139/ssrn.2117051.
- [33] P. Alves et al., “Strategic Talent Management: The Impact of Employer Branding on the Affective Commitment of Employees,” *Sustain.* 2020, Vol. 12, Page 9993, vol. 12, no. 23, p. 9993, Nov. 2020, doi: 10.3390/SU12239993.
- [34] P. J. Von Vultée, R. Axelsson, and B. Arnetz, “The impact of organisational settings on physician wellbeing,” *Int. J. Health Care Qual. Assur.*, vol. 20, no. 6, pp. 506–515, 2007, doi: 10.1108/09526860710819440.
- [35] P. Carayon and M. J. Smith, “Work organization and ergonomics,” *Appl. Ergon.*, vol. 31, no. 6, pp. 649–662, Dec. 2000, doi: 10.1016/S0003-6870(00)00040-5.
- [36] C. B. Gibson and S. G. Cohen, *Virtual Teams That Work: Creating Conditions for Virtual Team Effectiveness*. San Francisco: Jossey-Bass, 2003.
- [37] L. Feuer, “The challenge: Hire the right people for the right reasons,” *Case Manager*, vol. 11, no. 3, pp. 24–26, May 2000, doi: 10.1016/S1061-9259(00)80058-2.
- [38] D. Moher, F. Naudet, I. A. Cristea, F. Miedema, J. P. A. Ioannidis, and S. N. Goodman, “Assessing scientists for hiring, promotion, and tenure,” *PLOS Biol.*, vol. 16, no. 3, p. e2004089, Mar. 2018, doi: 10.1371/JOURNAL.PBIO.2004089.
- [39] J. E. Bono and T. A. Judge, “Personality and Transformational and Transactional Leadership: A Meta-Analysis,” *J. Appl. Psychol.*, vol. 89, no. 5, pp. 901–910, Oct. 2004, doi: 10.1037/0021-9010.89.5.901.
- [40] R. S. Dennis, L. Kinzler-Norheim, and M. Bocarnea, “Servant Leadership Theory,” *Servant Leadersh.*, pp. 169–179, 2010, doi: 10.1057/9780230299184\_14.
- [41] S. C. Davison, “Creating a High Performance International Team,” *J. Manag. Dev.*, vol. 13, no. 2, pp. 81–90, Mar. 1994, doi: 10.1108/02621719410050200/FULL/XML.
- [42] S. L. Jarvenpaa and D. E. Leidner, “Communication and Trust in Global Virtual Teams,” *Organ. Sci.*, vol. 10, no. 6, pp. 791–815, Dec. 1999, doi: 10.1287/orsc.10.6.791.
- [43] J. S. Gallego, I. Ortiz-Marcos, and J. Romero Ruiz, “Main challenges during project planning when working with virtual teams,” *Technol. Forecast. Soc. Change*, vol. 162, p. 120353, Jan. 2021, doi: 10.1016/J.TECHFORE.2020.120353.
- [44] J. A. Holton, “Building trust and collaboration in a virtual team,” *Team Perform. Manag. An Int. J.*, vol. 7, no. 3–4, pp. 36–47, Jun. 2001, doi: 10.1108/13527590110395621/FULL/XML.
- [45] H. K. Kim, “Work-Life Balance and Employees’ Performance: The Mediating Role of Affective Commitment,” *Glob. Bus. Manag. Res. An Int. J.*, vol. 6, no. 1, 2014.
- [46] J. P. Meyer, S. V. Paunonen, I. R. Gellatly, R. D. Goffin, and D. N. Jackson, “Organizational Commitment and Job Performance: It’s the Nature of the Commitment That Counts,” *J. Appl. Psychol.*, vol. 74, no. 1, pp. 152–156, 1989, doi: 10.1037/0021-9010.74.1.152.
- [47] J. P. Meyer and N. J. Allen, *Commitment in the workplace: theory, research, and application*. Sage Publications, 1997.
- [48] A. Westphal and M. Gmür, “Organisationales Commitment und seine Einflussfaktoren: Eine qualitative Metaanalyse,” *J. für Betriebswirtschaft* 2009 594, vol. 59, no. 4, pp. 201–229, Dec. 2009, doi: 10.1007/S11301-009-0054-X.
- [49] J. E. Mathieu and D. M. Zajac, “A Review and meta-analysis of the antecedents, correlates, and consequences of organizational commitment,” *Psychol. Bull.*, vol. 108, no. 2, pp. 171–194, 1990, doi: 10.1037/0033-2909.108.2.171.

# Fuzzy Image Enhancement Method based on a New Intensifier Operator

Libao Yang<sup>1</sup>, Suzelawati Zenian<sup>2,\*</sup>, Rozaimi Zakaria<sup>3</sup>

Faculty of Science and Natural Resources, Universiti Malaysia Sabah, Kota Kinabalu, Sabah, Malaysia<sup>1,2,3</sup>  
School of Mathematics and Information Technology, Xingtai University, Xingtai, Hebei, China<sup>1</sup>

**Abstract**—In recent years, fuzzy image enhancement methods have been widely applied in image enhancement, which generally consists of three steps: fuzzification, modify membership(using intensifier (INT) operator), and defuzzification. This paper proposed a new INT operator used in fuzzy image enhancement. The INT operator is adjustable for different test images. The image enhancement method is as follows, firstly, calculate the image threshold ( $T$ ) using the OTSU method. Secondly, calculate pivotal point  $p$  corresponding to  $T$ , and find the corresponding INT operator function. Finally, use the INT operator in fuzzy Image Enhancement. The INT operator is used multiple times in the image processing process to obtain multiple result images. Comparative experiments show that the proposed new INT operator has better image enhancement effect when INT operator is applied at the same number of times. On the other hand, more intermediate process result images can also be obtained through the proposed new INT operator. More result images can provide material resources for the subsequent image processing.

**Keywords**—Image enhancement; intensifier operator; threshold; pivotal point

## I. INTRODUCTION

Image enhancement's aim is to highlight useful information and remove useless information. The widely used image enhancement algorithms include gray transformation method [1]–[5], histogram equalization (HE) method [6]–[8], wavelet transform method [9]–[11] and basis algorithm Retinex method [12], [13] in color constancy theory. Gray transformation method is to directly apply the transformation function on the gray level to produce a new gray level. This method is relatively simple and easy to be implemented. HE method can increase image dynamic range and improve image contrast by making the probability density function of image gray level meet the form of approximately uniform distribution. The wavelet transform method divides the image into low frequency image and high frequency image, and enhances the different frequencies image to highlight the details of the image. Retinex method removes the influence of the illuminance component in the original image, and obtains the result image.

Image enhancement is widely used in many fields and many different types of images, such as infrared images, remote sensing images, underwater images, medical images and so on [14]–[17]. In recent years, fuzzy enhancement methods have been developed rapidly. Pal-King method [18] is the first proposed fuzzy image enhancement method, and has been applied in many fields until now. Generally, the fuzzy enhancement method consists of three steps: calculate

the image pixels' membership by fuzzification, adjust the membership using intensifier (INT) operator, and output new pixels' gray level by defuzzification. In the development of fuzzy image enhancement method, many functions of fuzzification, INT operator and defuzzification have been proposed. We introduce some common fuzzy enhancement methods. Li et al. [19] proposed a fast and reliable image enhancement technique based upon the fuzzy relaxation algorithm. Different orders of fuzzy membership functions and different rank statistics are attempted to improve the enhancement speed and quality, respectively. Hanmandlu et al. [20] used a Gaussian membership function to fuzzify the image information in spatial domain, and introduced a global INT operator which contains three parameters. Aiming at the membership cannot fill the interval  $[0, 1]$  and the pivotal point  $p$  of INT operator in Pal-King method is always the same ( $p \equiv 0.5$ ), Liu [21] proposed a fuzzification function based on tangent function and an INT operator with an adjustable parameter. Mahashwari et al. [22] proposed a defuzzification function. Hasikin et al. [23] proposed a new fuzzy intensity method to distinguish between the dark and bright regions. This method is computed by considering the average intensity and deviation of the intensity distribution of the image. The input image is enhanced using a power-law transformation. Singh et al. [24] proposed a new INT operator and a defuzzification method. Dawayet et al. [25] proposed a fuzzification and a defuzzification method. Yang et al. [26] proposed an INT operator based on cycloid arc length function. Fuzzy C-means clustering is also a good method for image enhancement [27]–[29]. More and more researchers are studying the fuzzy image enhancement method. The fuzzy image enhancement method has more and more unique advantages in the image enhancement field.

In existing methods, almost all fuzzification functions, INT operator functions and defuzzification functions are either piecewise functions or have a pivotal point  $p$  that cannot be adjusted ( $p \equiv 0.5$ ). In the realization of the algorithm (such as Matlab), the piecewise function needs to add the judgment statement in the coding, and use the judgment statement repeatedly in the calculation. This result will cause the computation time to increase. In view of pivotal point  $p \equiv 0.5$  and the possible shortcomings of piecewise function, this paper proposed an INT operator that is used in fuzzy image enhancement method. The INT operator function is made up by power function and has a variable parameter  $\lambda$ . For different test images, the parameter  $\lambda$  is determined prior to image enhancement. In Section 3, compare the image enhancement of the INT operator in [21] and the proposed INT operator using the same fuzzification function, and defuzzification function. The experimental results show that when INT operator process

\*Corresponding author

the same times, the proposed INT operator can achieve a better image enhancement, and even if the proposed INT operator is used many times, it still has a good enhancement effect.

## II. METHODOLOGY

Fuzzy image enhancement includes some steps: (1) using fuzzification to transform the image pixels' gray level to its membership, (2) modifying membership using intensifier (INT) operator, and (3) calculating pixels' new gray level by defuzzification. In this section, first, we introduce the common INT operators, and then propose a new INT operator.

### A. INT Operators

Although researchers have proposed many INT operators, there are two INT operators that appear most frequently in papers because they are simple to calculate and easy to manipulate. The first INT operator is a function made up of quadratic functions [18],

$$y = f(x) = \begin{cases} 2x^2, & 0 \leq x \leq 0.5, \\ 1 - 2(1-x)^2, & 0.5 < x \leq 1. \end{cases} \quad (1)$$

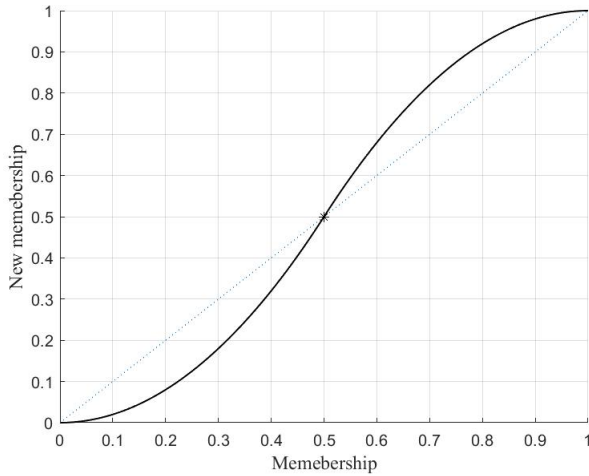


Fig. 1. The INT Operator  $y = f(x)$ .

Fig. 1 shows that the graph of INT operator  $y = f(x)$ . For all test images, the INT operator  $y = f(x)$ 's pivotal point (function piecewise point or inflection point)  $p$  is identically 0.5. This obviously has some limitations. In avoiding the shortcomings of INT operator  $y = f(x)$ , the improved INT operator is presented as follows [21]:

$$y = g(x) = \begin{cases} \frac{x^2}{p}, & 0 \leq x \leq p, \\ 1 - \frac{(1-x)^2}{1-p}, & p < x \leq 1. \end{cases} \quad (2)$$

The INT operator  $y = g(x)$  can adjust the point  $p$  according to the characteristics of the test image. Fig. 2 shows that the

INT operator  $y = g(x)$  can reduce the points which less than  $p$ , and enlarge the points which more than  $p$ . Both INT operator  $y = f(x)$  and INT operator  $y = g(x)$  are piecewise functions. The piecewise functions have some disadvantages in calculation, especially in the programming implementation (such as Matlab), judgment statements will be used many times in the program, which increases the amount of calculation. Now, we propose a new INT operator,

$$y = \varphi(x) = \frac{2x^\lambda}{1+x^\lambda}, \quad 0 \leq x \leq 1. \quad (3)$$

Where  $\lambda = 1 - \log_p(2-p)$ . The INT operator  $y = \varphi(x)$  is also can adjust the point  $p$  and its implementation procedure is more simplified. It does not need to use a judgment statement in the program. Fig. 2 shows that the graph of INT operator  $y = g(x)$  and INT operator  $y = \varphi(x)$ .

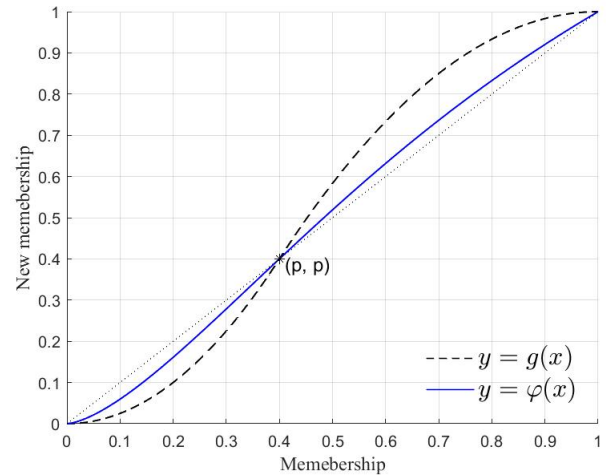


Fig. 2. An Example for of the INT Operators  $y = g(x)$  and  $y = \varphi(x)$  when  $p = 0.4$ .

### B. Fuzzy Contrast Enhancement

Image  $I = \{x_{ij} | i = 1, 2, 3, \dots, m, j = 1, 2, 3, \dots, n\}$ , where  $x_{ij}$  is gray level of the pixel in row  $i$  and column  $j$  of the image. To compare the effect of INT operators  $y = g(x)$  and  $y = \varphi(x)$ , we use the same fuzzification and defuzzification. Fuzzy contrast enhancement includes three steps as follow:

(1) Gray level fuzzification (Counting pixels' membership)

$$\mu_{ij} = \mu(x_{ij}) = \frac{x_{ij} - x_{min}}{x_{max} - x_{min}} \quad (4)$$

where,  $x_{max}(x_{min})$  is the maximum(minimum) pixel's gray level of the test image.

(2) Membership modification using INT operator:

$$\hat{y}_{ij} = I_t(u_{ij}) = I_1(I_{t-1}(u_{ij})) \quad t = 1, 2, 3, \dots$$

INT operator:  $I_1(u_{ij}) = g(u_{ij})$ ,

or INT operator:  $I_1(u_{ij}) = \varphi(x_{ij})$ .

In the Eq.(2) and Eq.(3),  $p = \mu(T)$ ,  $T$  is the test image threshold by the OTSU method.

(3)Pixels' new gray level by defuzzification

$$y_{ij} = (x_{max} - x_{min})\hat{y}_{ij} + x_{min}. \quad (5)$$

$I_{enh} = \{y_{ij} | i = 1, 2, 3, \dots, m, j = 1, 2, 3, \dots, n\}$  is the result image.

### III. EXPERIMENTAL RESULTS AND ANALYSIS

Structural similarity(SSIM) is commonly used to evaluate the image enhancement effect. This section also use the SSIM value as an objective evaluation criterion for image enhancement.

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)}. \quad (6)$$

In equation (6),  $\mu_x$  is the mean of  $x$ ,  $\mu_y$  is the mean of  $y$ ,  $\sigma_x^2$  is the variance of  $x$ ,  $\sigma_y^2$  is the variance of  $y$ ,  $\sigma_{xy}$  is the covariance of  $x$  and  $y$ .  $c_1$  and  $c_2$  are constants [30].



Fig. 3. Test Images.

Fig. 3 shows the test images. INT operator  $y = g(x)$  and INT operator  $y = \varphi(x)$  processed test images for 2 and 8 times, respectively.

Fig. 4 shows that at the same number of processing times, the image processed by INT operator  $y = \varphi(x)$  is closer to the original image, and more intermediate process images can be obtained. When the INT operators are used the same number of times, the INT operator  $y = \varphi(x)$  has a higher SSIM value. It achieves a better image enhancement effect.

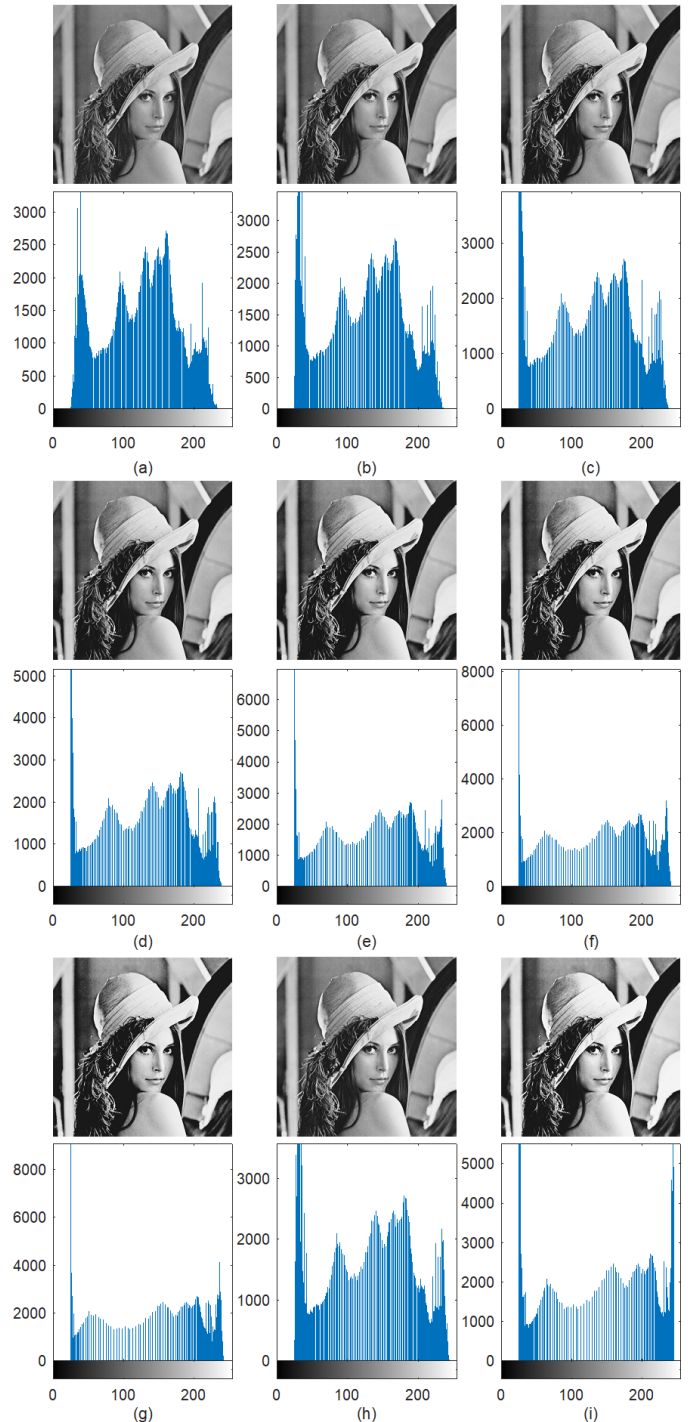


Fig. 4. Enhancement Results of the Lena and its Histogram. (a)-(g). Processed by Proposed INT Operator  $y = \varphi(x)$  Times:1-7; (h)-(i). Processed by Method with INT Operator  $y = g(x)$  Times:1-2.



TABLE I. STRUCTURAL SIMILARITY (SSIM) TEST RESULTS

Test images		Lena	Couple	Fishing Boat	Peppers
processed by INT operator $y = g(x)$	t = 1	0.9341	0.9000	0.9016	0.8871
	t = 2	0.7963	0.7125	0.7462	0.6793
processed by proposed INT operator $y = \varphi(x)$	t = 1	0.9916	0.9882	0.9863	0.9781
	t = 2	0.9681	0.9538	0.9485	0.9244
	t = 3	0.9325	0.9044	0.8987	0.8615
	t = 4	0.8911	0.8477	0.8514	0.7968
	t = 5	0.8485	0.7924	0.8117	0.7330
	t = 6	0.8090	0.7422	0.7785	0.6720
	t = 7	0.7725	<b>0.6990</b>	0.7503	0.6181
	t = 8	0.7398	0.6615	0.7238	0.5736

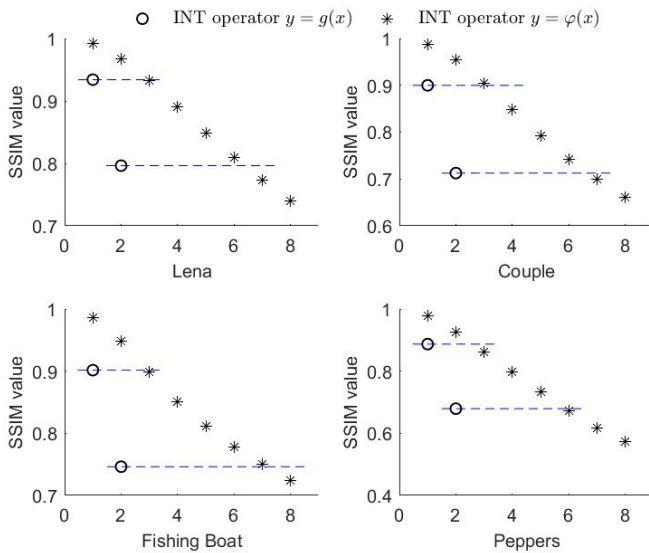


Fig. 5. The SSIM Values when the Test Images are Processed by Fuzzy Image Enhancement Method using INT Operators t Times.

In Table I, for the convenience of expression, denoted  $S(\text{image}, \text{INT operator}, t)$  means that the SSIM value of image and the image which is processed t times by INT operator. For example,  $S(\text{Couple}, y = \varphi(x), 7) = 0.6990$ . Table I shows  $S(\text{Couple}, y = \varphi(x), 1) > S(\text{Couple}, y = \varphi(x), 2) > S(\text{Couple}, y = \varphi(x), 3) > S(\text{Couple}, y = g(x), 1)$ . This means that when INT operator  $y = g(x)$  processed 1 time, the INT operator  $y = \varphi(x)$  can processed 3 times. All of the 3 result images' SSIM values are higher than the result image processed by the INT operator  $y = g(x)$  1 time. For other test images, a similar phenomenon exists. Fig. 5 shows this more clearly. For example, Fig. 5 (Peppers) shows that  $S(\text{Peppers}, y = \varphi(x), 1) > S(\text{Peppers}, y = \varphi(x), 2) > S(\text{Peppers}, y = g(x), 1) > S(\text{Peppers}, y = \varphi(x), 3) > S(\text{Peppers}, y = \varphi(x), 4) > S(\text{Peppers}, y = \varphi(x), 5) > S(\text{Peppers}, y = g(x), 2) > S(\text{Peppers}, y = \varphi(x), 6) > S(\text{Peppers}, y = \varphi(x), 7) > S(\text{Peppers}, y = \varphi(x), 8)$ . Fig. 5 is a visual representation of Table I.

#### IV. CONCLUSION

This paper proposed a new intensifier(INT) operator used in fuzzy image enhancement. It has the following advantages. First, compared to the INT operator which is made up by a

piecewise function, the proposed INT operator has an advantage in program implementation that it does not require writing judgment statements. Second, due to the range of change of the proposed INT operator is small (see Fig. 2), the result image with higher SSIM value can be obtained after the INT operator is multiple applied. This means that it achieves a better image enhancement effect. It can produce more process images. More images of the process can provide material resources for the subsequent image processing, and could help in other areas of research. In future work, we will investigate whether it can be used in other fields.

#### ACKNOWLEDGMENT

The authors would like to express their appreciation and gratitude to the Research Management Centre, Universiti Malaysia Sabah for granting this research study under Skim UMSGreat (GUG0540-2/2020).

#### REFERENCES

- [1] A. Raji, A. Thaibaoui, E. Petit, P. Bunel, and G. Mimoun, *A gray-level transformation-based method for image enhancement*, Pattern Recognition Letters. 1998, vol. 19 no.13, p. 1207-12.
- [2] H. Gao, W. Zeng, and J. Chen, *An improved gray-scale transformation method for pseudo-color image enhancement*, Computer Optics. 2019, vol. 43 no.1, p. 78-82.
- [3] Y. R. K. Y. Zhang and C. Feng, *Image enhancement algorithm based on quadratic function and its implementation with fpga*, Modern Electronics Technique. 2020, vol. 43 no.8, p. 72-76.81.
- [4] B. Zhang, D. Xiao, L. Wang, S. Bai, and L. Yang, *Efficient compressed sensing based image coding by using gray transformation*, 2021, ArXiv preprint arXiv:2102.01272.
- [5] L. Yang, S. Zenian, and R. Zakaria, *An image enhancement method based on a S-sharp function and pixel neighborhood information*, Borneo Science. 2021, vol. 42 no.1, p. 18-24.
- [6] M. Kaur, J. Kaur, and J. Kaur, *Survey of contrast enhancement techniques based on histogram equalization*, Journal of Advanced Computer Science and Applications. 2011, vol. 2 no.7, p. 1-5.
- [7] S. C. F. Lin, C. Y. Wong, M. A. Rahman, G. Jiang, S. Liu, N. Kwok, H. Shi, Y. H. Yu, and T. Wu, *Image enhancement using the averaging histogram equalization (AVHEQ) approach for contrast improvement and brightness preservation*, Computers and Electrical Engineering. 2015, vol. 46, p. 356-370.
- [8] G. Raju and M. S. Nair, *A fast and efficient color image enhancement method based on fuzzy-logic and histogram*, International Journal of electronics and communications. 2014, vol. 68 no.3, p. 237-243.
- [9] Y. Yang, Z. Su, and L. Sun, *Medical image enhancement algorithm based on wavelet transform*, Electronics letters. 2010, vol. 46 no.2, p. 120-121.
- [10] C. Jung, Q. Yang, T. Sun, Q. Fu, and H. Song, *Low light image enhancement with dual-tree complex wavelet transform*, Journal of Visual Communication and Image Representation. 2017, vol. 42, p. 28-36.
- [11] M. X. Yang, G. J. Tang, X. H. Liu, L. Q. Wang, Z. G. Cui, and S. H. Luo, *Low-light image enhancement based on Retinex theory and dual-tree complex wavelet transform*, Optoelectronics Letters. 2018, vol. 14 no.6, p. 470-475.
- [12] A. Zotin, *Fast algorithm of image enhancement based on multi-scale retinex*, Procedia Computer Science. 2018, vol. 131, p. 6-14.
- [13] N. Hassan, S. Ullah, N. Bhatti, H. Mahmood, and M. Zia, *The Retinex based improved underwater image enhancement*, Multimedia Tools and Applications. 2021, vol. 80 no.2, p. 1839-57.
- [14] N. Sadic, E. Hassan, S. El-Rabaie, S. El-dolil, M. I. Dessoky, and F. El-samie, *Enhancement Technique of Infrared Images*, Menoufia Journal of Electronic Engineering Research. 2021, vol. 30 no.1, p. 58-64.
- [15] Z. Zhu, Y. Luo, H. Wei, Y. Li, G. Qi, N. Mazur, Y. Li, and P. Li, *Atmospheric light estimation based remote sensing image dehazing*, Remote Sensing. 2021, vol. 13 no.13, p. 24-32.



- [16] M. J. Islam, Y. Xia, and J. Sattar, *Fast underwater image enhancement for improved visual perception*, IEEE Robotics and Automation Letters. 2020, vol. 5 no.2, p. 3227-3234.
- [17] N. Salem, H. Malik, and A. Shams, *Medical image enhancement based on histogram algorithms*, Procedia Computer Science. 2019, vol. 163, p. 300-311.
- [18] S. K. Pal and R. A. King, *Image enhancement using fuzzy sets*, Electronics Letters. 1980, vol. 16 no.9, p. 376-378.
- [19] H. Li and H. S. Yang, *Fast and reliable image enhancement using fuzzy relaxation technique*, IEEE transactions on systems, man, and cybernetics. 1989, vol. 19 no.5, p. 1276-1281.
- [20] M. Hanmandlu and D. Jha, *An optimal fuzzy system for color image enhancement*, IEEE Transactions on image processing. 2006, vol. 15 no. 10, p. 2956-2966.
- [21] X. Liu, *An improved image enhancement algorithm based on fuzzy set*, Physics Procedia. 2012, vol.33, p. 790-797.
- [22] T. Mahashwari and A. Asthana, *Image enhancement using fuzzy technique*, International Journal of Research in Engineering Science and Technology. 2013, vol. 2 no. 2, p. 1-4.
- [23] K. Hasikin and N. A. M. Isa, *Fuzzy image enhancement for low contrast and non-uniform illumination images*, In 2013 IEEE International Conference on Signal and Image Processing Applications, October 2013, p. 275-280.
- [24] J. K. Singh and G. Shrivastava, *Fuzzy Logic Based Contrast Image Enhancement Technique*, International Journal of Research in Computer and Communication Technology. 2013, vol. 2 no. 12, p. 1448-1453.
- [25] H. G. Daway, E. G. Daway, and H. H. Kareem, *Colour image enhancement by fuzzy logic based on sigmoid membership function*, International Journal of Intelligent Engineering and Systems. 2020, vol. 13 no. 5, p. 238-246.
- [26] L. Yang, S. Zenian, and R. Zakaria, *Fuzzy image enhancement based on algebraic function and cycloid arc length*, 2021 IEEE International Conference on Artificial Intelligence in Engineering and Technology (IICAIET), September 2021, p. 1-4.
- [27] R. P. Sharma and S. Dey, *Two-stage quality adaptive fingerprint image enhancement using Fuzzy C-means clustering based fingerprint quality analysis*, Image and Vision Computing. 2019, vol. 83, p. 1-16.
- [28] M. M. Riaz, A. Ghafoor, and V. Sreeram, *Fuzzy C-means and principal component analysis based GPR image enhancement*, 2013 IEEE Radar Conference (RadarCon13), April 2013, p. 1-4.
- [29] L. Yang, S. Zenian, and R. Zakaria, *Image enhancement method based on an improved Fuzzy C-means clustering*, International Journal of Advanced Computer Science and Applications. 2022, vol. 13 no. 8, p. 855-859.
- [30] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, *Image quality assessment: from error visibility to structural similarity*, IEEE transactions on image processing. 2004, vol. 13 no. 4, p. 600-612.

# Cooperative Multi-Robot Hierarchical Reinforcement Learning

Gembong Edhi Setyawan<sup>1</sup>, Pitoyo Hartono<sup>2</sup>, Hideyuki Sawada<sup>3</sup>

Department of Applied Physics, School of Advanced Science and Engineering<sup>1,3</sup>

Waseda University, 3-4-1 Okubo, Shinjuku-ku, Tokyo 169-8555, Japan<sup>1,3</sup>

School of Engineering, Chukyo University<sup>2</sup>

101-2 Yagoto Honmachi, Showa-ku, Nagoya, Aichi, 466-8666 Japan<sup>2</sup>

**Abstract**—Recent advances in multi-robot deep reinforcement learning have made it possible to perform efficient exploration in problem space, but it remains a significant challenge in many complex domains. To alleviate this problem, a hierarchical approach has been designed in which agents can operate at many levels to complete tasks more efficiently. This paper proposes a novel technique called Multi-Agent Hierarchical Deep Deterministic Policy Gradient that combines the benefits of multiple robot systems with the hierarchical system used in Deep Reinforcement Learning. Here, agents acquire the ability to decompose a problem into simpler subproblems with varying time scales. Furthermore, this study develops a framework to formulate tasks into multiple levels. The upper levels function to learn policies for defining lower levels' subgoals, whereas the lowest level depicts robot's learning policies for primitive actions in the real environment. The proposed method is implemented and validated in a modified Multiple Particle Environment (MPE) scenario.

**Keywords**—Multi-robot system; hierarchical deep reinforcement learning; path-finding; task decomposition

## I. INTRODUCTION

Multi-Robot System (MRS) research has attracted significant attention recently due to its advantages over single robots. The benefits include (1) the decrease in time and the improvement in problem-solving efficiency due to the task decomposition, (2) an increase in problem-solving reliability, robustness, and resiliencies as the failures of single robots can be [1]–[3]. Using MRS, numerous prospective applications have been developed in which robots must be able to compete and cooperate, such as formation coordination [4], hide and seek [5], exploration and search [6]–[8], object transportation [9], disaster detection [10], communication networks [11], [12], etc. This study focuses on using MRS technology for exploration and search missions in unknown environments, where robots must collaborate to find the optimal path.

As shown in [13], [14], Reinforcement Learning (RL) is currently extensively employed as a robot learning algorithm that can automatically handle exploration and search problems in unknown environments, both in simulation and physical environments. Multiple robots undertake search and exploration operations in large and complex environments, such as in [15]–[17]. The objective of MRS is to distribute tasks between many robots to increase efficiency. Nowadays, applying RL is an important and challenging subject in MRS,

where robots must learn and adapt based on their individual strategies and collective behavior. Multi-Agent Deep Deterministic Policy Gradient (MADDPG) algorithm was proposed as an advancement of RL that may be used for multi-robot learning in which robots can collaborate to solve problems in unknown environments [18].

RL in complex environments is always challenging, hence it serves as the primary motivation for the proposed approach. This paper proposes Multi-Agent Hierarchical Deep Deterministic Policy Gradient (MH-DDPG) as an extension of MADDPG for solving search and exploration problems with many robots in finding the optimal path for each robot in various environmental complexities. Here, hierarchical learning is adopted to learn in complex environments efficiently [13].

The primary contribution of this paper is as follows: (1) MH-DDPG presents a framework that enables multiple robots to learn by sharing environmental information collectively. (2) MH-DDPG proposes a hierarchical learning strategy by assigning different abstraction levels to the problem space, where higher abstraction learnings supply the learning subgoal for the lower ones, which consequently execute automatic task decomposition.

The structure of this article is as follows: Section II examines the previous research and theories relevant to the proposed method. In Section III, the technical context and theory are explained. The proposed MH-DDPG is discussed in detail in Section IV. Sections V and VI present the conducted experiments and the validation of the results. Finally, section VII provides conclusions and potential future works.

## II. RELATED WORKS

Reinforcement Learning (RL) [19] is increasingly used as learning method to address complicated problems like games [20], [21], and robotics [22], [23]. Q-Learning [24], SARSA [25], and Temporal Differences (TD) [26] are RL algorithms that are predominantly applied to single agents within the Markov Decision Process (MDP) mathematical modeling framework [27].

Traditional RL faces various problems when attempting to solve real-world problems. This issue is known as the "curse of dimensionality", in which data grows exponentially, and computations become costly. Deep Reinforcement Learning (DRL) was introduced to solve this issue, for example, in [20].

---

This work was supported by JSPS Grants-in-Aid for Scientific Research on Innovative Areas (Research in a proposed research area) 18H05473 and 18H05895.

DRL models complex functions using the advantages of Neural Networks (NN), for example, Deep Q-network (DQN) [20], [28]. However, implementing DQN for robots with many degrees of freedom in action or continuous action space remains challenging. Policy Gradient (PG) [29] has been proposed to solve these issues; nevertheless, PG has its challenges, mainly that it requires extensive training. An actor-critic algorithm [25] combining DQN and PG's benefits was proposed to mitigate these problems. In the actor-critic algorithm, there are two networks: the actor-network, which employs a policy gradient to create optimal policies, and the critic-network, which contains DQN to evaluate the actor's policies. Deep Deterministic Policy Gradient (DDPG) [30] is an actor-critic [31] based algorithm that has been proposed to solve problems in continuous or high-dimensional action spaces.

Multi-Robot System (MRS) is currently attracting much interest because it can solve complex problems that are prohibitively difficult for single-agent systems. Although several attempts have been made to apply the prior algorithm in MRS, it is still not as successful as in single-agent systems. The recent development of MRS operates within the stochastic (Markov) games framework to model mathematical decision-making [32]. The difficulty of developing MRS is that the policies of one robot will impact a non-stationary environment generated by the policies of another agent. Recently, Multi-agent Deep Deterministic Policy Gradient (MADDPG) [18] was presented as a solution to this issue. For each robot, there is a DDPG in the MADDPG. The MADDPG is able to manage the problems of competitive as well as collaborative multi-agent systems.

One attempt to alleviate the learning difficulty is to design a mechanism for hierarchical learning. Hierarchical learning involves decomposing large and complex problems into more manageable subproblems. However, most algorithms need to fix the sub-problems that may hinder problem-solving flexibility. The Hierarchical Reinforcement Learning (HRL) approach of a single robot with discrete action has been proposed in [33], [34]; however, the subgoals need to be assigned manually, while [35] has been designed to be able to find subgoals automatically. Hierarchical Deep Reinforcement Learning (HDRL), [36], [37] proposed systems that can operate with continuous robot action. Algorithms in [33]–[37] work well with a single robot but are unsuitable for multi-robot implementation. Studies in [38], [39] presented a multi-agent hierarchy system with DRL for learning subgoals/skills at higher levels. However, the higher-level environment was manually defined using these approaches.

It is known that MADDPG can handle collaborative multi-robot system problems in a simple environment. However, an algorithm that can improve the learning performance of multi-robot systems is required for more complex environments. In this study, we propose MH-DDPG by developing MADDPG to perform under a hierarchical system. The proposed method is expected to be able to automatically discover subgoals, allowing it to perform better in a high-complexity environment.

### III. TECHNICAL PRELIMINARIES

This section highlights the theoretical basis for developing a hierarchical MADDPG method for cooperative multi-agent learning in handling complex problems.

#### A. Markov Decision Process and Reinforcement Learning

The decision-making process of a single robot is often based on Markov Decision Process (MDP). In RL, at each time step  $t$ , the robot perceives a state  $S_t$  from the environment's set of states  $\mathcal{S}$  ( $S_t \in \mathcal{S}$ ) and the robot selects an action  $A_t$  from the set of actions  $\mathcal{A}$  that may be executed in the state  $S_t$  ( $A_t \in \mathcal{A}$ ). Here, the next state is decided based on transition probability  $\mathcal{P}$  under the learned policy  $\pi$ , as shown in (1). The transition brings the robot to the next state  $S_{t+1}$  and gives reward  $R_{t+1}$ .

$$\pi(s', r|s, a) = \mathcal{P}(S_{t+1} = s', R_{t+1} = r | S_t = s, A_t = a) \quad (1)$$

where  $s \in \mathcal{S}$  and  $s' \in \mathcal{S}$  are particular states that occur at time  $t$  and  $t+1$ ,  $a \in \mathcal{A}$  is the action taken by the robot at time  $t$  ( $a \in \mathcal{A}$ ), and ( $r \in \mathcal{R}$ ) is the reward received by the robot at time  $t+1$ .

Typically, the Bellman Equation is used to optimize two functions in RL: the state ( $V^*$ ) and state-action ( $Q^*$ ) functions, for which the optimal equation is as follows:

$$V^*(s) = \max_{\pi} R(s, a) + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}_{ss'}^a V^*(s') \quad (2)$$

$$Q^*(s, a) = R(s, a) + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}_{ss'}^a Q^*(s', a') \quad (3)$$

Here, the discount factor ( $0 \leq \gamma \leq 1$ ) is used to express the significance of the future reward value.

#### B. Stochastic (Markov) Games

In contrast to MDP, which is utilized for a single robot, stochastic games are proposed as a mathematical framework for modeling decision-making in Multi-robot Reinforcement Learning. Stochastic games consist of  $N$  robots, a set of states containing the state of all robots ( $\mathcal{S}$ ), a set of actions  $\mathcal{A}$  from all robots ( $A = A_1 \times A_2 \times \dots \times A_N$ ), and a set of state transitions ( $T$ ) which are the transition probability ( $\mathcal{P}$ ) from the current state to the next state for a robot based on the actions taken by all robots ( $T: \mathcal{S} \times A_1 \times A_2 \times \dots \times A_N \rightarrow \mathcal{P}(\mathcal{S})$ ). The reward obtained by a robot depends on the actions taken by all robots ( $R: \mathcal{S} \times A_1 \times A_2 \times \dots \times A_N \rightarrow \mathcal{R}$ ). The reward function for each robot can be used to classify the type of games. For example, all robots share the same reward function if they play cooperatively. In contrast, when the robots play competitively, one robot aims to maximize the reward while the other attempts to minimize it. In stochastic games, the state value function ( $V$ ) could be written as follows:

$$V_{i,\pi}(s) = \mathbb{E}[\sum_{k=0}^{\infty} \gamma^k r_{i,t+k+1} | S_{i,t} = s_i] \quad (4)$$

#### C. Q-Learning and Deep Q-Network

Q-Learning is a RL mechanism based on Q-function. Similar to (3), the Q-function is used to compute the expected reward based on the action done by the robot in its current state. The optimum policy is found by maximizing the value of the Q-function. The Q-value is learned iteratively, as follows:

$$Q(s, a) \leftarrow Q(s, a) + \alpha \left( R + \gamma \left( \max_{a'} Q(s', a') - Q(s, a) \right) \right) \quad (5)$$

where  $0 \leq \alpha \leq 1$  denotes the learning rate.

The Q-value for all potential state-action combinations is stored in a Q-table, and thus high computational resources are required in a large number of states and large action space. The necessity for these costly computation resources can be alleviated by replacing the Q-table with a neural network for estimating the Q-value as in Deep Q-Network (DQN).

Q-Network and target Network are the two neural networks that constitute a DQN. The Q-network is used to train robots to predict the optimal Q-value, while the target network is used to forecast the next state based on the sample data and the optimal Q-value from all potential actions in the next state. In addition, DQN has a component called Experience Replay (ER) that stores and generates training data for Q-Network.

The optimal policy for DQN is determined by minimizing the Loss function in (6).

$$L(Q) = \mathbb{E}_{s,a,r,s'} \left[ \left( R(s, a) + \gamma \max_{a'} Q^*(s', a' | \bar{\theta}) - Q(s, a | \theta) \right)^2 \right] \quad (6)$$

where  $\bar{\theta}$  and  $\theta$  are the parameters for the target network and the Q-network.

#### D. Policy Gradient

Policy Gradient (PG) is used to enhance DQN's performance in generating optimum policies. In DQN, the robot chooses an action with the maximum Q-value, while in PG, the agent selects an action stochastically according to the probability distribution generated in the output layer.

The PG consists of a neural network known as a policy network, which predicts the probability distribution of actions given the current state. Here, the optimal policy is determined by maximizing the objective function defined as follows:

$$J(\theta) = \sum_{s \in \mathcal{S}} d_{\pi}(s) \sum_{a \in \mathcal{A}} \pi_{\theta}(a, s) Q_{\pi}(s, a) \quad (7)$$

where  $d_{\pi}(s)$  is the deterministic distribution of the states on  $\pi$ . Here, the objective function  $J(\theta)$  can be maximized by adjusting the parameter  $\theta$  by gradient  $\nabla_{\theta} J(\theta)$  as follows:

$$\nabla_{\theta} J(\theta) = \mathbb{E}_{s \sim d_{\pi}, a \sim \pi_{\theta}} [\nabla_{\theta} \log \pi_{\theta}(a, s) Q_{\pi}(s, a)] \quad (8)$$

#### E. Deterministic Policy Gradient and Deep Deterministic Policy Gradient

The policy function in PG is always modeled as a stochastic probability distribution of the agent's actions given the current state. The Deterministic Policy Gradient (DPG) has been proposed to model policy as a deterministic decision by the agent in the current state. The objective function in DPG can be written as follows:

$$J(\theta) = \mathbb{E}_{s \sim \rho_{\pi}} [R(s, \pi_{\theta}(s))] \quad (9)$$

where  $\rho_{\pi}$  is discounted state distribution. The gradient of the objective function in DPG can be written as follows:

$$J(\theta) = \mathbb{E}_{s \sim \rho_{\pi}} [\nabla_{\theta} \pi_{\theta}(s) \nabla_{\theta} Q_{\pi}(s, a) |_{a=\pi_{\theta}(s)}] \quad (10)$$

Deep Deterministic Policy Gradient (DDPG) is an actor-critic algorithm that combines DQN and DPG. DQN is for the actors that operate in discrete action space, while DPG is for the critics that work in continuous action space.

#### F. Multi-Agent Deep Deterministic Policy Gradient

Multi-Agent Deep Deterministic Policy Gradient (MADDPG) is an expansion of DDPG that adopt an actor-critic algorithm as its fundamental structure. The MADDPG contains multiple robots, each with its neural networks for the actors and the critics, while DDPG only uses a single robot. Similar to DDPG, the actors in MADDPG receive input from the robot's local observations and produce executable action recommendations for the robot. However, in contrast to the critical network in DDPG, the input of critics in MADDPG does come from not only the robot's local observations and actions but also other robots' observations and actions. The critic's output is the Q-value, which is used to evaluate the actor's actions by considering other robots' observations and acts. The network of agents may therefore learn both cooperative and competitive strategies.

### IV. MULTI-HIERARCHIES OF MULTI-AGENT DEEP DETERMINISTIC POLICY GRADIENT

This study proposes Multi-Agent Hierarchical Deep Deterministic Policy Gradient (MH-DDPG) as a new approach that enables learning robots to decompose complex tasks into more manageable subtasks at different time scales. Here, the robots train to learn several levels of policy, each of which has a specific task for the agents to do in parallel.

#### A. Architecture

MH-DDPG trains robots to hierarchically learn policies based on the architecture shown in Fig. 1. Here, MH-DDPG is comprised of DDPG and experience replay (ER). The number of DDPG and ER depends on the number of agents and hierarchy. For example, suppose  $N$  and  $K$  indicate the number of agents and hierarchies, respectively. Consequently, there are  $N \times K$  DDPG in MH-DDPG. Furthermore, the number of ER will equal the number of hierarchies,  $K$ . The bottom level represents the physical environment in which the robots physically operate. While the higher level, robots are presented by their abstraction.

Formally, the MH-DDPG with  $N$  agents and  $K$  levels are defined by the set of state  $\mathcal{S}$ ; the set of joint action  $\mathcal{A} = \cup_{i=1}^N \mathcal{A}_{i,t}^{\ell}$  and the set of joint observations  $\mathcal{O} = \cup_{i=1}^N \mathcal{O}_i^{\ell}$ , where  $\mathcal{A}_{i,t}^{\ell}$  and  $\mathcal{O}_i^{\ell}$  are actions  $A_i$  and observation  $O_i$  for each agent  $i$  at level  $\ell$  and time  $t$ . Each agent will optimize their respective policy at every level to estimate the transition probability  $\mathcal{P}$  for selecting an action at time  $t$ , such that  $\pi = \cup_{i=1}^N \pi_{i,t}^{\ell}$ , where  $\ell$  is the hierarchy level and  $0 \leq \ell \leq K - 1$ .

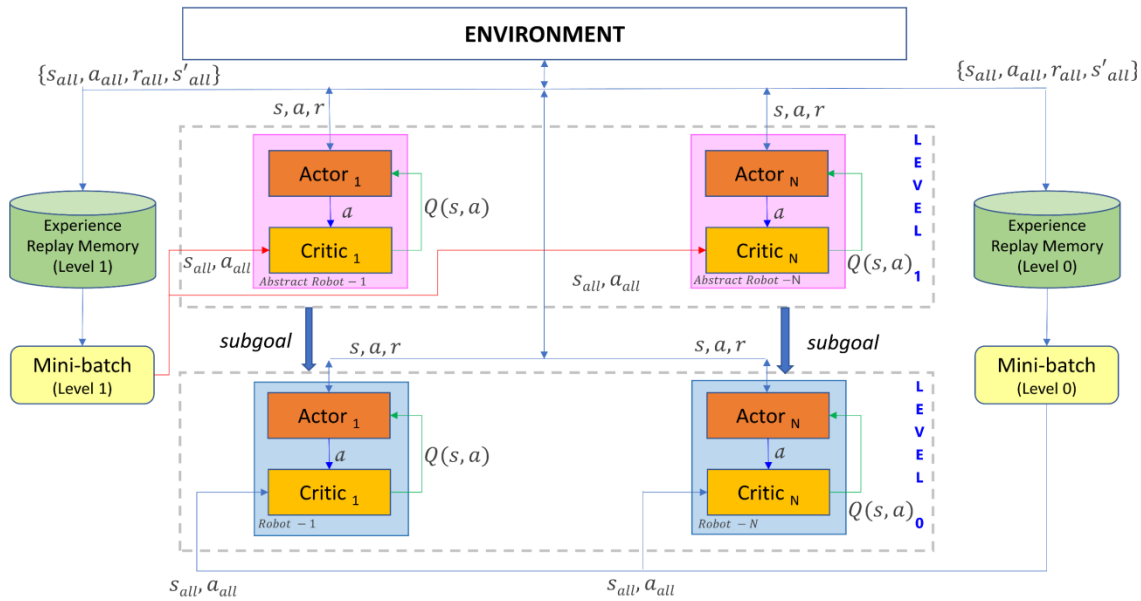


Fig. 1. Architecture of MH-MADDPG.

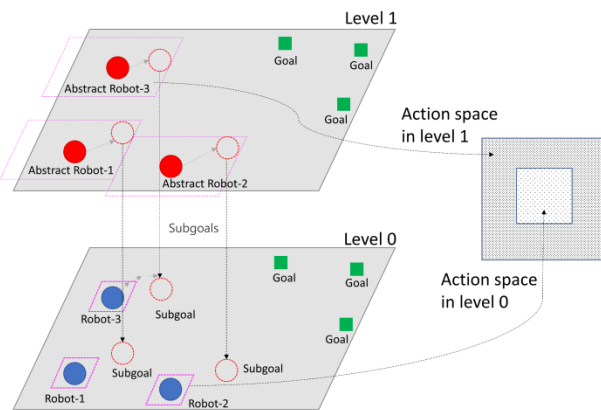


Fig. 2. Illustration Environment with Two Hierarchies.

For illustrating the dynamics of MH-DDPG multi-agent particle environment (MPE) environments will be utilized [40]. One of the MPE scenarios, "simple spread" has been modified here. For example, Fig. 2 depicts the problem that MH-DDPG must address. Simple\_spread is an environment with  $N$  robots and  $M$  goals (landmarks). Robots are expected to cooperate to accomplish a common objective while avoiding collisions with one another. There are three robots ( $N=3$ ), three goals ( $M=3$ ), and two hierarchies ( $K=2$ ). First, the real problem of the environment is illustrated at level 0, where a blue circle represents the actual robots, and a green rectangle represents the goals. Then, at level 1, an abstraction of level 0, the robots are referred to as abstract robots and symbolized by a red circle. Abstract robots at level 1 possess actions with more capabilities than those of actual robots at level 0. For instance, the actual robots at level 0 have a maximum velocity of 1 pixel per second, while the robot at level 1 is set with a maximum velocity of 10 pixels per second. As seen on the right of Fig. 2, the robot at level 1 has a greater range of distances than the actual robot for each action taken at each step.

The abstract robots are predicted to learn faster than the actual robots in achieving goals since they are less constrained than actual robots (for example, more quickly and with no obstacles). However, remember that abstract robots are only imaginative robots with no capacity to execute physical actions. The task of the abstract robot at level 1 is to learn how to accomplish the main goal best, while at level 0, the task is to learn how to achieve the subgoal optimally. MH-DDPG implicitly assigns different objectives for each level, in which the robots' objective at level 1 is to learn to achieve goals optimally, while the robots' work at level 0 is to learn to achieve subgoals optimally. The subgoal at level 0 is automatically determined from the higher level, which happens when the abstract robot chooses the action  $A_{i,t}^1$  based on the policy  $\pi_{i,t}^1$  in the current state  $S_{i,t}^1$  at time  $t$ . The robot will be transitioned to the next state  $S_{i,t+1}^1$  and will receive a reward  $R_{i,t+1}^1$ . Then, the learning shifts to the bottom level, and the next state at level 0 becomes a subgoal for the actual robots. In addition, robots at level 0 engage in learning to achieve these subgoals. When the robots get  $S_{i,t}^0$  at level 0, the agent will pick the action  $A_{i,t}^0$  based on the policy of  $\pi_{i,t}^0$ . The robots then transition to  $S_{i,t+1}^0$  and is rewarded with  $R_{i,t+1}^0$ . The learning process at level 0 will continue until the terminal criteria are satisfied. A terminal condition is defined by manually setting the maximum number of steps at level 0. If the terminal requirements are satisfied, learning returns to level 1 to execute the next step at the top level.

### B. Learning Dynamic

In MH-DDPG, the multiple robots learn in parallel at all levels. The process of robot learning will start at the top level and flows downward. Robot learning aims to provide optimum policies for each robot at all levels. According to the RL concept that the optimal policy is acquired by maximizing the rewards received by each robot. The reward obtained in the

future by each robot  $i$  at level  $k$  could be expressed by the following V-function:

$$V_{i,\pi}^k(s) = \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r_{i,t+1}^k | S_{i,t}^k = s] \quad (11)$$

$$V_{i,\pi}^k(s) = \sum_{a \in A} \pi_i^k(a|s) (R_i^k(s, a) + \gamma \sum_{s' \in S} \mathcal{P}_{ss'}^a V_{i,\pi}^k(s')) \quad (12)$$

where  $r_{i,t+1}^k$  is the reward earned by agent  $i$  at level  $k$  at  $t + 1$ , and  $\pi_i^k(a|s)$  is the agent policy. Here, the joint action is designed to make the robot's policy dependent on individual policies and joint policies. The definition of the Q-function is as follows:

$$Q_{i,\pi}^k(s, a) = \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r_{i,t+1}^k | S_{i,t+1}^k = s, A_{i,t}^k = a] \quad (13)$$

$$Q_{i,\pi}^k(s, a) = R_i^k(s, a) + \gamma \sum_{s' \in S} \mathcal{P}_{ss'}^a \sum_{a' \in A} \pi_i^k(a'|s') Q_{i,\pi}^k(s', a') \quad (14)$$

The optimal policy is determined by maximizing the value of all actions. According to the Bellman optimality equation, the optimal V-value ( $V^*$ ) and Q-value ( $Q^*$ ) could be written as follows:

$$V_i^k(s) = \max_{\pi_i} R_i^k(s, a) + \gamma \sum_{s' \in S} \mathcal{P}_{ss'}^a V_i^k(s') \quad (15)$$

$$Q_i^k(s, a) = R_i^k(s, a) + \gamma \sum_{s' \in S} \mathcal{P}_{ss'}^a Q_i^k(s', a') \quad (16)$$

If the environment consists of  $N$  agents and  $K$  levels, then the policy set  $\pi = \{\pi_1^k, \pi_2^k, \dots, \pi_N^k\}$  that is parameterized by  $\theta = \{\theta_1^k, \theta_2^k, \dots, \theta_N^k\}$ , where  $1 \leq i \leq N$  and  $0 \leq k \leq K - 1$ . Then, the gradient of the expected return for each agent  $i$  at level  $k$  could be expressed as follows:

$$J(\theta_i^k) = \mathbb{E}_{x,a \sim D} [R(s, \pi_i^k(s))] \quad (17)$$

$$\begin{aligned} \nabla_{\theta_i^k} J(\theta_i^k) &= \mathbb{E}_{x,a \sim D} \\ & \left[ \nabla_{\theta_i^k} \log \pi_i^k(a_i^k | \mathcal{O}_i^k) Q_{i,\pi}^k(x^k, a_1^k, \dots, a_N^k) \Big|_{a_i^k = \pi_i^k(\mathcal{O}_i^k)} \right] \end{aligned} \quad (18)$$

where  $Q_{i,\pi}^k(x, a_1^k, \dots, a_N^k)$  is the centralized Q-function at level  $k$  that accepts as input all agent actions at level  $k$ ,  $a_1^k, \dots, a_N^k$ , and observation  $x$  at level  $k$  of all agents,  $x = (\mathcal{O}_1^k, \dots, \mathcal{O}_N^k)$ , with the output being the Q-value for each agent  $i$  at level  $k$ . Experience Replay buffer  $\mathcal{D}$  contains  $(x, x', a_1^k, \dots, a_N^k, r_1^k, \dots, r_N^k)$  where  $x'$  is the next state obtained after the agent took action while in state  $x$ . The centralized Q-function  $Q_{i,\pi}^k$  will be updated by minimizing the following loss function:

$$\mathcal{L}(\theta_i^k) = \mathbb{E}_{x,a,r,x'} [(Q_{i,\pi}^k(x, a_1^k, \dots, a_N^k) - y)^2] \quad (19)$$

$$y = r_i^k + \gamma Q_{i,\pi',k}^k(x', a_1^{k'}, \dots, a_N^{k'}) \Big|_{a_j^{k'} = \pi_j^{k'}(\mathcal{O}_j^{k'})} \quad (20)$$

where  $\pi^{k'} = \{\pi_{\theta_1^{k'}}, \dots, \pi_{\theta_N^{k'}}\}$  is the set of target policy with delayed parameter  $\theta_i^{k'}$  at each level  $k$ . As the Q-function  $Q_{i,\pi}^k$  for each agent  $i$  is learned independently at all levels, the reward may be determined arbitrarily based on the issue. The algorithm of MH-DDPG is shown in algorithm 1.

#### Algorithm 1. MH-MADDPG

---

```

1  Initialize: Actor-critic evaluation and target networks for each agent,
2      number of levels K, maximum step H, Replay buffer
3  For episode = 1 to max-episode, do
4      For each agent i, set initial states (S) and goals (G) for each agent
5      Train (K-1, S, G)
6  End for
7
8  Function Train( $k$  ::level, S ::state, G ::goal)
9       $s \leftarrow S_{i,t}^k \leftarrow S, g^k \leftarrow G$  (initial, t=0)
10
11  For t = 1 to H do
12      For each agent n: select action ( $a_i$ ) where  $a_i \leftarrow A_{i,t}^k$  based on  $\pi_{i,t}^k$ 
13      Execute actions  $a = (a_1, \dots, a_N)$  and observe reward r and new
14      state  $s'$ 
15      Store (s, a, r,  $s'$ ) in replay buffer  $\mathcal{D}$ 
16      If  $k > 0$ :
17           $g^{k-1} \leftarrow s'$ 
18           $Train(k-1, s, g^{k-1})$ 
19      End If
20      For agent i = 1 to N in level  $k$  do
21          Sample random minibatch of S samples (s, a, r,  $s'$ ) from  $\mathcal{D}^k$ 
22          Set  $y = r_i^k + \gamma Q_{i,\pi}^k(s', a_1^k, \dots, a_N^k) \Big|_{a_i^k = \pi_i^k(s')}$ 
23          Update critic by minimizing the loss  $\mathcal{L}(\theta_i^k) = \frac{1}{S} \sum (y -$ 
24               $Q_{i,\pi}^k(s, a_i, \dots, a_N))^2$ 
25          Update actor using:
26               $\nabla_{\theta_i^k} J =$ 
27               $\frac{1}{S} \sum \nabla_{\theta_i^k} \pi_i^k(\mathcal{O}_i) \nabla_{a_i} Q_{i,\pi}^k(s, a_i, \dots, a_N) \Big|_{a_i = \pi_i^k(\mathcal{O}_i)}$ 
28      End for
29      Update target network parameters for each agent i in level  $k$ :
30           $\theta_i^{k'} \leftarrow \alpha \theta_i^k + (1 - \alpha) \theta_i^{k'}$ 
31  End Function

```

---

#### C. State, Observation, and Action Space

Consider an environment with  $N$  robots and  $M$  goals. Robots and goals have a physical entity represented by  $X$ . Based on the original MPE,  $X$  is a two-dimensional object characterized by its position and velocity. Furthermore, the state contains polar coordinates that are utilized to identify the robot's relative position to the goals and other robots. An environment with  $N$  robots and  $M$  goals corresponds to a state space with  $N \times M$  polar coordinates of robots to the goals ( $d_{1,\dots,N \times M}^G$ ) and  $N-1$  polar coordinates to other robots ( $d_{1,\dots,N-1}^A$ ). However, it should be noted that the goals of the bottom level are the subgoals produced at the upper level.

Based on the preceding discussion, the state space  $\mathcal{S}$  is a mixture of each level state space:  $\mathcal{S} = \bigcup_{k=0}^{K-1} S^k$ , where  $S^k = \{X_{1,\dots,N}, d_{1,\dots,N \times M}^G, d_{1,\dots,N-1}^A\}$ .

Then each agent can only observe their own state of the entire state, called observation. The observation space of each agent at each level  $k$  is  $O_i^k(S) = \{X_i, d_{i,1,\dots,M}^G, d_{i,1,\dots,N-1}^A\}$ , where  $i$  indicates the  $i^{th}$  robot.

At  $k = 0$ , the output layer of the actor networks generates five outputs between 0 and 1 in which each one is associated with a particular action. The five outputs are denoted by  $u_n$ ,  $u_l$ ,  $u_r$ ,  $u_d$ , and  $u_u$  for no action, move left, right, down, and up, respectively. At  $k > 0$ , if the action of the abstract robot should have more capabilities than the actual robot, the range  $u$  is increased multiplied by the sensitivity rate, therefore  $u^k = u^0 x \mu$ , where  $1 \leq k \leq K - 1$  and  $\mu$  is the sensitivity with a value more than 1. The sensitivity of the upper level must be larger than the sensitivity of the lower level.

#### D. Reward Design

The reward is designed to correspond to the learning objectives of the robot. The distance between objects determines the reward design. Suppose that the positions of two object types, A and B, in two dimensions are known. A and B respectively add up to  $N$  and  $M$ , therefore  $A_i = (x_1^i, y_1^i)$  and  $B_j = (x_2^j, y_2^j)$ , where  $1 \leq i \leq N$  and  $1 \leq j \leq M$ . The following formula may be used to compute the total distance between two types of objects:

$$d(A, B) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \quad (21)$$

$$d(A_{1,\dots,N}, B_{1,\dots,M}) = \sum_{i=1}^N \sum_{j=1}^M \sqrt{(x_2^i - x_1^j)^2 + (y_2^i - y_1^j)^2} \quad (22)$$

The design of the reward differs between the bottom and upper levels. The term for the rewards at each level is explained as follows:

1) *The goal/subgoal reward*: Designed to encourage agents to achieve the Goal/Subgoal. This reward is available at all levels. This reward is utilized at the highest level to promote the abstract robot to accomplish the main goal and at the lowest level to help the robot reach the subgoal. Reward calculations will be based on the distance between all robots and goals using (22) and (23).

$$R(\text{Agent}_{i,\dots,N}, \mathcal{G}_{j,\dots,M}) = -d(A_{1,\dots,N}, B_{1,\dots,M}) \quad (23)$$

where  $A = \text{Robot}$  dan  $B = G$  (goal/subgoal).

2) *Robot relative to other robots Reward*: for avoiding collisions between robots. This reward is only used at the lowest level because the abstract robot is unable to detect other robots. This reward term is calculated as follows:

$$R_c(A, B) = \begin{cases} -1; & \text{if } d(A, B) \leq A_{size} + B_{size} \\ 0; & \text{if } d(A, B) > A_{size} + B_{size} \end{cases} \quad (24)$$

where  $d(A, B)$  is the distance between two robots ( $A$  and  $B$ ) that can be calculated by (22) with  $i, j=I$ .

3) *Obstacle reward*: Aims to encourage robots to avoid obstacles. Due to the abstract robot's inability to detect obstacles, this reward is only applied at the lowest level. Similar to other robot rewards, the robot must compute the distance between itself and the obstacle to get reward.

$$R_c(A, B) = \begin{cases} -10; & \text{if } d(A, O) \leq A_{size} + O_{size} \\ 0; & \text{if } d(A, O) > A_{size} + O_{size} \end{cases} \quad (25)$$

where  $d(A, O)$  is the distance between robot  $A$  and obstacle  $O$  that can be calculated by (22) with  $i, j=I$ .

#### E. Neural Network Models

Each robot at each level of the MH-DDPG consists of actor and critic networks, structures of which are shown in Fig. 3. Local observations are the inputs for the actor-network, while robot actions represent the output. Therefore, the critic-network uses the observations and actions of all robots as inputs and Q-value as outputs. The Q-value is then used as the training basis for the actor networks. Every network employs the ADAM optimizer with a learning rate ( $\alpha$ ) and a discount factor ( $\gamma$ ).

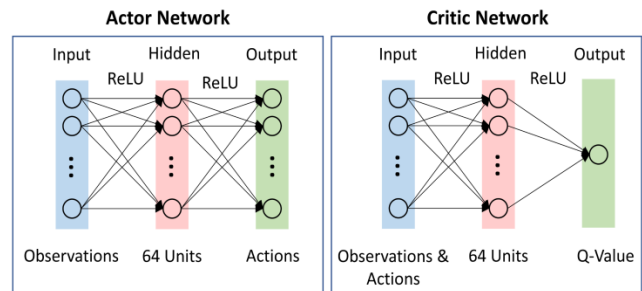


Fig. 3. Neural Network Model.

#### V. EXPERIMENTAL ENVIRONMENT

We conducted experiments for comparing the DDPG, MADDPG, and MH-DDPG algorithms under the parameters listed in Table I. The experimental environment is set with three robots and three goals, with some obstacles to increase the environment's complexity. Fig. 4 depicts the environment for testing the proposed algorithm. The experiments were conducted on three types of environments with various complexities: low-complexity {Fig. 4(a)}, mid-complexity {Fig. 4(b)}, and high-complexity {Fig. 4(c)}.

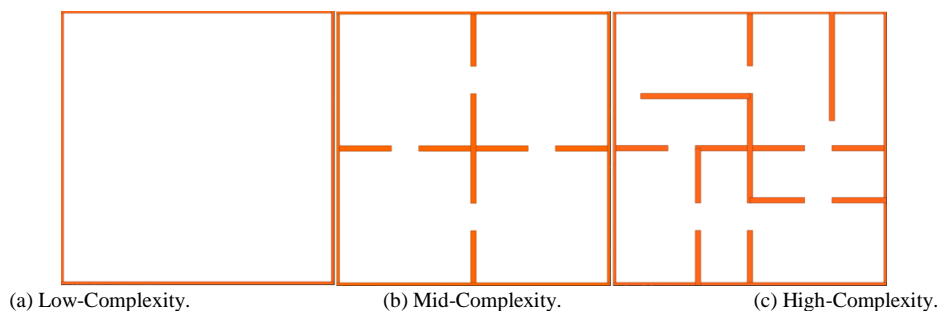


Fig. 4. Illustration of the Experimental Environment.



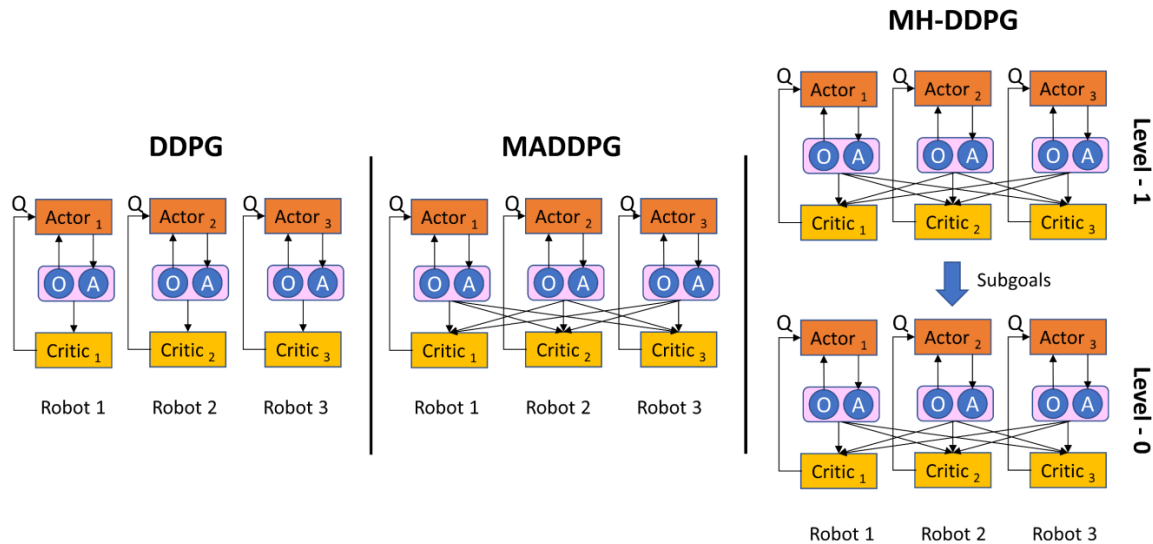


Fig. 5. Block Diagram Comparison for DDPG, MADDPG, and MH-DDPG Employing 3 Robots. Q, O, and a Represent the Q-Value, Observation, and Action, Respectively.

TABLE I. EXPERIMENT-SPECIFIC PARAMETERS FOR DDPG, MADDPG, AND MH-DDPG

Parameters	DDPG	MADDPG	MH-DDPG
<b>Specific Parameters</b>			
The number of robots (N)	3	3	3
The number of levels (K)	-	-	2
The number of actors	3	3	6
The number of critics	3	3	6
The number of ERs	1	1	2
Sensitivity	5	5	5 (level 0), 30 (level 1)
<b>Actor and Critic Networks Parameters</b>			
Number of hidden layers	1	1	1
Number of hidden units	64	64	64
Activation Function	ReLU	ReLU	ReLU
Input Actor Network	Current Observations	Current Observations	Current Observations
Output Actor Network	Action	Action	Action
Input Critic Network	Current Observation and Action	Current Observation and Action	Current Observation and Action
Output Critic Network	Q-value	Q-Value	Q-value
<b>Training parameters</b>			
Optimizer	ADAM	ADAM	ADAM
Learning rate ( $\alpha$ )	1e-2	1e-2	1e-2
Discount factor ( $\gamma$ )	0.97	0.97	0.97
Replay buffer size	$10^6$	$10^6$	$10^6$
Minibatch size	1256	1256	1256

Fig. 5 compares the block diagrams of DDPG, MADDPG, and MH-DDPG, illustrating how the algorithm determines the parameter values for the specific parameters given in Table I, except for sensitivity. The values for the sensitivity and the training parameters are empirically determined. From the experiments, the determination of the training parameters in different environments demonstrates that the algorithms are not excessively sensitive to the chosen parameters.

Particularly in MH-DDPG, the designed environment can decompose into K levels. As a preliminary step in the proposed algorithm, this research performs a two-level investigation (K=2). Where the bottom level ( $k = 0$ ) is the real environment used for actual robots learning, and the top level ( $k = 1$ ) is the abstract environment used for abstract robot learning. In the environments, the robots must collaborate to accomplish the predetermined goals. The robots' mission will be accomplished if the robots can discover the optimal path for reaching all the goals.

## VI. RESULTS

The experiments compare the proposed algorithm against MADDPG and DDPG. The first step is assessing the robots learning performance based on the rewards obtained throughout the learning process.

Fig. 6 depicts the learning curve based on the robot's average reward in each episode. In this experiment, there were 150000 episodes in each environment. The average reward the robot obtains in a low-complexity environment is greater than in mid-complexity and high-complexity environments. A greater average reward indicates that robots in simple environments perform better than in other environments. A low-complexity environment without obstacles makes it easier for robots to reach their goals.

From Fig. 6, it can be observed that MH-DDPG converges faster to the maximum rewards and produces larger reward in each episode than DDPG and MADDPG, indicating that the robots that were trained using MH-DDPG reaches the goals

faster. Fig. 6 also indicates that the superiority of MH-DDPG over DDPG and MADDPG is consistent in complex environments. In a high-complexity environment, the graph also reveals that the average reward value is unstable for DDPG and MADDPG, whereas MH-DDPG remains robust.

Fig. 7 depicts the robot's behavior during the learning process compared to MADDPG and DDPG in a mid-complexity and high-complexity environment at  $t=0, 10,$  and  $40$ . The blue circle indicates the actual robot, the green rectangle represents the goals, and the red circle represents the abstract robot found exclusively on MH-DDPG. The abstract robots will generate a subgoal for the lower level. At  $t=0$ , the robot begins its first step of learning. Here, the locations of the abstract robot are identical to the positions of the actual robots.

At  $t=10$ , robots are learning to achieve all goals. Here, on MH-DDPG, a red circle indicates the presence of an abstract robot. At each instant  $t$ , the abstract robot generates a subgoal for the actual robot. In MH-DDPG, Actual robots will first learn to cover subgoals, but in DDPG and MADDPG, robots

will learn to cover main goals straight away since there are no subgoals. Abstract robots that cannot detect obstacles might occasionally be located in the same area as the obstacle, as shown in the high-complexity environment at time  $t=10$ . This condition is sometimes harmful to the actual robot when it learns to achieve the subgoal since the actual robot cannot reach the subgoal properly, which consequently decreases the associated state-action values.

Finally, at  $t=40$ , the final step of learning in a single episode occurs. In DDPG, it is evident that the robots have difficulty cooperating to reach the goal in both mid-complexity and high-complexity environments, which is also consistent with the successful rates shown in Fig. 8. In the mid-complexity environment, the robot tends to go toward one of the goals. Therefore the targeted objectives to cover all goals are often not achieved. This is because the robots work independently and do not share information. In high-complexity, obstacles tend to hinder the robot's ability to achieve the subgoals.

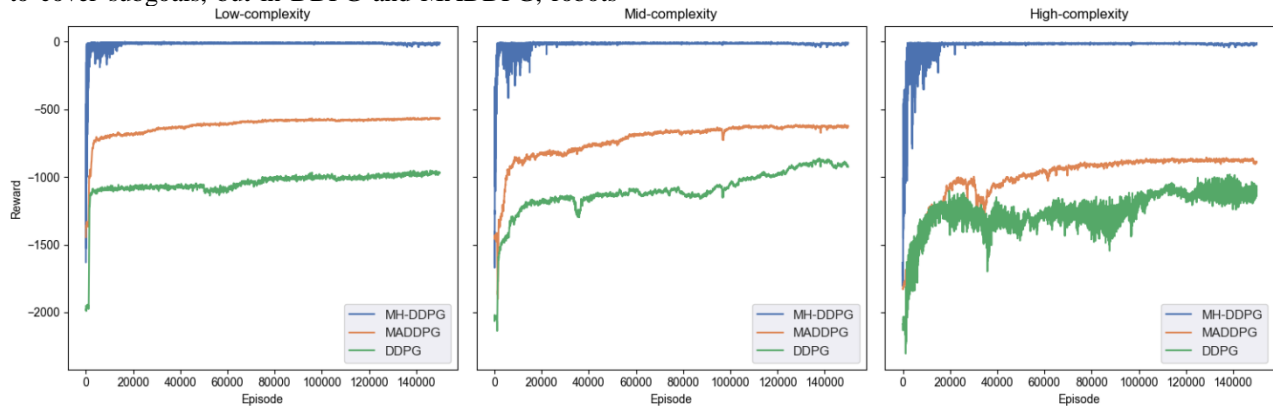


Fig. 6. Reward Graph.

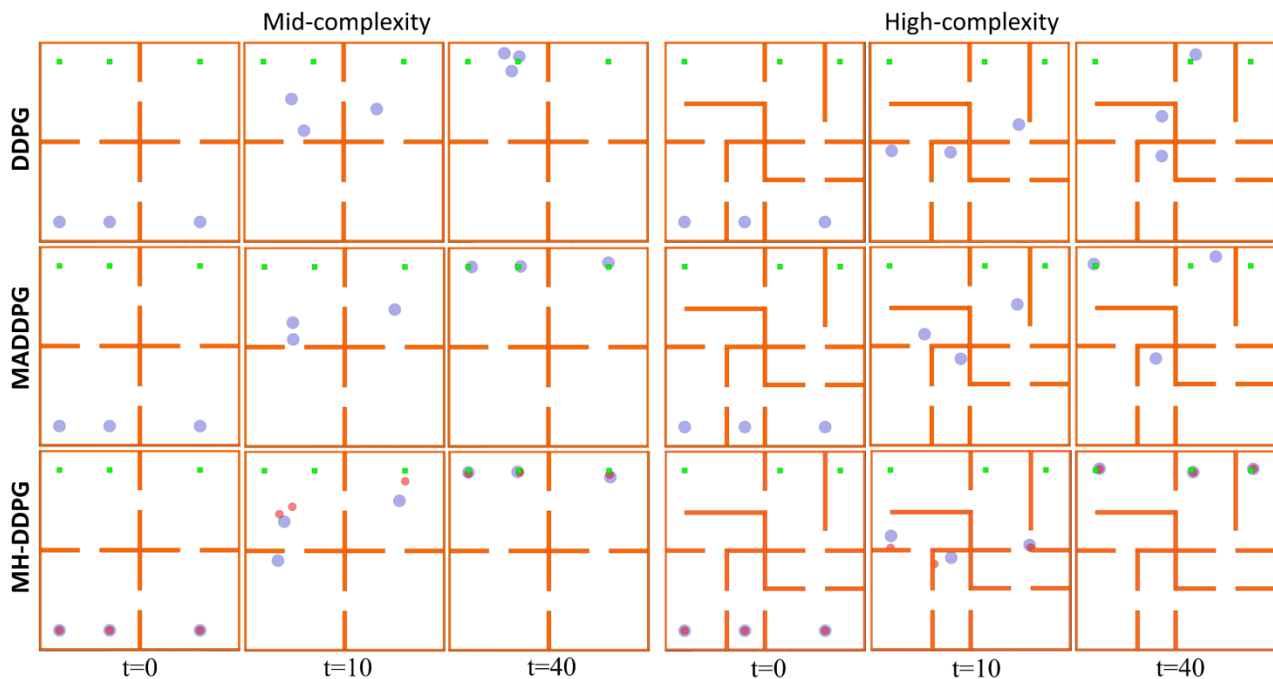


Fig. 7. Comparison between DDPG, MADDPG, and MH-DDPG on the Mid-complexity and High-complexity Environment.

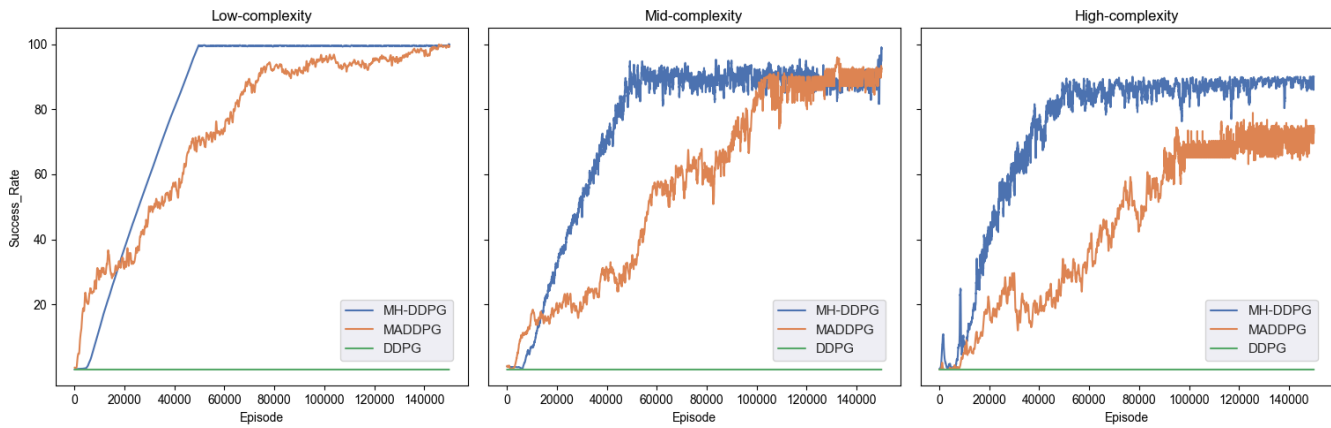


Fig. 8. Success Rate.

From Fig. 7, for MADDPG, it can be observed that the robots can work collectively to achieve goals in a mid-complexity environment, while in high-complexity environments, due to the complex configurations of the obstacles, the robots faced difficulty in the early phase of the learning process as apparent from the fluctuating graph, but gradually stabilize as the learning progresses. In MH-DDPG, subgoals increase the robot's problem-solving ability as they are guided by intermediate objectives that consequently constrained the problem space.

Fig. 8 depicts the average success rate for the respective algorithm in each environment during the learning process. In all environments, the robots with MADDPG and MH-DDPG were able to cooperate and learn policies to achieve their goals, while the robots with DDPG failed to do so. The failure of the robots in DDPG is due to the lack of information sharing between robots; hence, the robots only learn independently and may repeat the failure of other robots. In the low-complexity environment, the average success rates for MADDPG and MH-DDPG are 74.18% and 82.15%, respectively. As the complexity of the environment increases, the average success rate for MADDPG and MH-DDPG decreases, as seen from the graphs for mid and high-complexity environments. MADDPG and MH-DDPG had respective success rates of 56.08% and 73.18% in a mid-complexity environment, while in a high-complexity environment, these success rates were 44.18% and 72.99%, respectively. The results show that MH-DDPG has a greater success rate than MADDPG. This indicates that decomposing the problem environment into many levels is advantageous for maximizing robot learning performance.

## VII. CONCLUSION

MH-DDPG is proposed as a novel framework for multi-robot learning with hierarchical Deep Reinforcement Learning. Here, the robots collectively learn by sharing information about state-action values from their individual runs. In addition, the proposed MH-DDPG provides a mechanism for creating multi-level abstraction, in which higher-level abstraction space allow the robots to execute a kind of "image training" where they may virtually explore the problem space without considering the physical constraints in real-world space. The virtual experiment in abstract space allows the robot to discover the real robots' intermediate goals rapidly. The intermediate goals

helps to limit the exploration for the real robots, thus alleviating the curse of dimensionality.

Through some empirical experiments, it can be observed that the proposed MH-DDPG outperforms DDPG and MADDPG in learning efficiency and success rate.

The weakness of the MH-DDPG is that an abstract robot at higher levels is incapable of detecting obstacles. Hence non-realistic subgoals are sometimes produced. This is the cost that needs to be paid for removing the physical constraints in the abstract space. In this preliminary experiment, the abstract robots are given higher speed but are constrained by their inability to detect obstacles, but it does not have to be so. In the following study, experiments will be conducted with various constrained conditions at the higher abstraction levels.

Immediate future research topics include investigating the effect of the number of levels and the number of robots in MH-DDPG. In addition, implementing the proposed learning method into physical robots is also of interest.

## REFERENCES

- [1] Z. Yan, N. Jouandeau and A. A. Cherif, "A survey and analysis of multi-robot coordination," *Int. J. Adv. Robot. Syst.*, vol. 10, pp. 1–18, 2013.
- [2] J. Song and S. Gupta, "CARE: Cooperative Autonomy for Resilience and Efficiency of robot teams for complete coverage of unknown environments under robot failures," *Auton. Robots*, vol. 44, no. 3–4, pp. 647–671, 2020.
- [3] Y. Rizk, M. Awad and E. W. Tunstel, "Cooperative heterogeneous multi-robot systems: A survey," *ACM Comput. Surv.*, vol. 52, no. 2, 2019.
- [4] Z. A. Ali, Z. Han and R. J. Masood, "Collective motion and self-organization of a swarm of uavs: A cluster-based architecture," *Sensors*, vol. 21, no. 11, pp. 1–19, 2021.
- [5] B. Baker, I. Kanitscheider, T. Markov, Y. Wu, G. Powel, B. McGrew and I. Mordatch, "Emergent tool use from multi-agent autocurricula," 2020.
- [6] M. Bakhshpour, M. Jabbari Ghadi and F. Namdari, "Swarm robotics search & rescue: A novel artificial intelligence-inspired optimization approach," *Appl. Soft Comput. J.*, vol. 57, pp. 708–726, 2017.
- [7] J. P. Queralta, J. Taipalmaa, B. C. Pullinen, V. K. Sarker, T. N. Gia, H. Tenhunen, M. Gabbouj, J. Raitoharju and T. Westerlund, "Collaborative multi-robot search and rescue: Planning, coordination, perception, and active vision," *IEEE Access*, vol. 8, pp. 191617–191643, 2020.
- [8] D. S. Drew, "Multi-Agent Systems for Search and Rescue Applications," *Curr. Robot. Reports*, vol. 2, no. 2, pp. 189–200, 2021.

- [9] L. Hawley and W. Suleiman, "Control framework for cooperative object transportation by two humanoid robots," *Rob. Auton. Syst.*, vol. 115, pp. 1–16, 2019.
- [10] X. Zhou, W. Wang, T. Wang, Y. Lei and F. Zhong, "Bayesian Reinforcement Learning for Multi-Robot Decentralized Patrolling in Uncertain Environments," *IEEE Trans. Veh. Technol.*, vol. 68, no. 12, pp. 11691–11703, 2019.
- [11] N. Naderializadeh, J. J. Sydir, M. Simsek and H. Nikopour, "Resource Management in Wireless Networks via Multi-Agent Deep Reinforcement Learning," *IEEE Trans. Wirel. Commun.*, vol. 20, no. 6, pp. 3507–3523, 2021.
- [12] J. Sheng, X. Wang, B. Jin, J. Yan, W. Li, T.-H. Chang, J. Wang and H. Zha, "Learning structured communication for multi-agent reinforcement learning," *Auton. Agent. Multi. Agent. Syst.*, vol. 36, no. 2, 2022.
- [13] F. Niroui, K. Zhang, Z. Kashino and G. Nejat, "Deep Reinforcement Learning Robot for Search and Rescue Applications: Exploration in Unknown Cluttered Environments," *IEEE Robot. Autom. Lett.*, vol. 4, no. 2, pp. 610–617, 2019.
- [14] S. M. Sombolistan, A. Rasooli and S. Khodaygan, "Optimal path-planning for mobile robots to find a hidden target in an unknown environment based on machine learning," *J. Ambient Intell. Humaniz. Comput.*, vol. 10, no. 5, pp. 1841–1850, 2019.
- [15] J. Hu, H. Niu, J. Carrasco, B. Lennox and F. Arvin, "Voronoi-Based Multi-Robot Autonomous Exploration in Unknown Environments via Deep Reinforcement Learning," *IEEE Trans. Veh. Technol.*, vol. 69, no. 12, pp. 14413–14423, 2020.
- [16] R. J. Alitappeh and K. Jeddisaravi, "Multi-robot exploration in task allocation problem," *Appl. Intell.*, vol. 52, no. 2, pp. 2189–2211, 2022.
- [17] T. Fan, P. Long, W. Liu and J. Pan, "Distributed multi-robot collision avoidance via deep reinforcement learning for navigation in complex scenarios," *Int. J. Rob. Res.*, vol. 39, no. 7, pp. 856–892, 2020.
- [18] R. Lowe, Y. Wu, A. Tamar, J. Harb, P. Abbeel and I. Mordatch, "Multi-agent actor-critic for mixed cooperative-competitive environments," *Adv. Neural Inf. Process. Syst.*, vol. 2017-Decem, pp. 6380–6391, 2017.
- [19] R. S. Sutton and A. G. Barto, *Reinforcement Learning: an Introduction*, 2nd ed. London, England: The MIT Press Cambridge, Massachusetts, 1998.
- [20] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg and D. Hassabis, "Human-level Control through Deep Reinforcement Learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.
- [21] X. Wang, J. Song, P. Qi, P. Peng, Z. Tang, W. Zhang, W. Li, X. Pi, J. He, C. Gao, H. Long and Q. Yuan, "SCC: an efficient deep reinforcement learning agent mastering the game of StarCraft II," 2021.
- [22] Y. Chebotar, K. Hausman, Y. Lu, T. Xiao, D. Kalashnikov, J. Varley, A. Irpan, B. Eysenbach, R. Julian, C. Finn and S. Levine, "Actionable Models: Unsupervised Offline Reinforcement Learning of Robotic Skills," 2021.
- [23] M. Dalal, D. Pathak and R. Salakhutdinov, "Accelerating Robotic Reinforcement Learning via Parameterized Action Primitives," in 35th Conference on Neural Information Processing Systems (NeurIPS 2021), 2021, no. NeurIPS.
- [24] C. J. C. H. Watkins and P. Dayan, "Q-Learning," *Mach. Learn.*, vol. 8, pp. 279–292, 1992.
- [25] G. A. Rummery and M. Niranjan, *On-Line Q-Learning Using Connectionist Systems*. Technical Report CUED/F-INFENG/TR 166, Cambridge University Engineering Department, 1994.
- [26] R. S. Sutton, "Learning to Predict by the Methods of Temporal Differences," *Mach. Learn.*, vol. 3, no. 1, pp. 9–44, 1988.
- [27] M. L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. New Jersey: John Wiley & Sons, Inc., 1994.
- [28] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra and M. Riedmiller, "Playing Atari with Deep Reinforcement Learning," vol. arXiv prep, 2013.
- [29] R. S. Sutton, D. McAllester, S. Singh and Y. Mansour, "Policy Gradient Methods for Reinforcement Learning with Function Approximation," in *Proceedings of the 12th International Conference on Neural Information Processing Systems*, 1999, pp. 1057–1063.
- [30] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver and D. Wierstra, "Continuous control with deep reinforcement learning," 4th Int. Conf. Learn. Represent. ICLR 2016 - Conf. Track Proc., 2016.
- [31] V. R. Konda and J. N. Tsitsiklis, "Actor-critic algorithms," *Adv. Neural Inf. Process. Syst.*, pp. 1008–1014, 2000.
- [32] M. L. Littman, "Markov games as a framework for multi-agent reinforcement learning," *Mach. Learn. Proc.* 1994, pp. 157–163, 1994.
- [33] R. S. Sutton, D. Precup and S. Singh, "Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning Richard," *Artif. Intell.*, vol. 112, no. 1, pp. 181–211, 1998.
- [34] A. Bai and S. Russell, "Efficient reinforcement learning with hierarchies of machines by leveraging internal transitions," in *IJCAI International Joint Conference on Artificial Intelligence*, 2017, vol. 0, pp. 1418–1424. [Online].
- [35] G. E. Setyawan, H. Sawada and P. Hartono, "Combinations of Micro-Macro States and Subgoals Discovery in Hierarchical Reinforcement Learning for Path Finding," *Int. J. Innov. Comput. Inf. Control*, vol. 18, no. 2, pp. 447–462, 2022.
- [36] A. S. Vezhnevets, S. Osindero, T. Schaul, N. Heess, M. Jaderberg, D. Silver and K. Kavukcuoglu, "FeUdal networks for hierarchical reinforcement learning," 34th Int. Conf. Mach. Learn. ICML 2017, vol. 7, pp. 5409–5418, 2017.
- [37] Z. Yang, K. Merrick, S. Member, L. Jin, H. A. Abbass and S. Member, "Hierarchical Deep Reinforcement Learning for Continuous Action Control," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 29, no. 11, pp. 5174–5184, 2018.
- [38] Z. Liang, J. Cao, W. Lin, J. Chen and H. Xu, "Hierarchical Deep Reinforcement Learning for Multi-robot Cooperation in Partially Observable Environment," *Proc. - 2021 IEEE 3rd Int. Conf. Cogn. Mach. Intell. CogMI 2021*, pp. 272–281, 2021.
- [39] J. Yang, I. Borovikov and H. Zha, "Hierarchical cooperative multi-agent reinforcement learning with skill discovery," *Proc. Int. Jt. Conf. Auton. Agents Multiagent Syst. AAMAS*, vol. 2020-May, pp. 1566–1574, 2020.
- [40] H. Clark and S. Brennan, "Emergence of grounded compositional language in multi-agent populations," 32nd AAAI Conf. Artif. Intell. AAAI 2018, pp. 1495–1502, 2018.

# Differential Privacy Technology of Big Data Information Security based on ACA-DMLP

Yubiao Han\*, Lei Wang, Dianhong He  
Shandong Information Technology Industry Development  
Research Institute, Jinan Shandong 250014, China

**Abstract**—Cloud computing and artificial intelligence have a deeper and closer connection with daily life. To ensure information security, most companies or individuals choose to pay a simple fee to store a large amount of data on cloud servers and hand over a large number of complex calculations of machine learning to cloud servers. To eliminate the security risks of data stored in the cloud and ensure that private data is not leaked, this paper proposes a collusion-resistant distributed machine learning scheme. Through homomorphic encryption algorithm and differential privacy algorithm, the security of data and model in machine learning framework is guaranteed. The distributed machine learning framework is adopted to reduce the data computing time and improve the data training efficiency. The simulation results show that the computational efficiency is improved while the user privacy security is guaranteed. The accuracy of model training is not reduced due to the improvement of privacy data security and computational efficiency. Through this study, we can further propose effective measures for the privacy protection of outsourced data and the data integrity of machine learning, which is of great significance to the security research of cloud intelligent big data.

**Keywords**—Big data; cloud computing; information security; distributed machine learning; differential privacy algorithms

## I. INTRODUCTION

The core technology of artificial intelligence is machine learning. Machine learning is mainly through the analysis of a large number of data, statistics, calculations, and other operations, from which to learn experience, build models, and step by step improve the accuracy of model training. In practice, machine learning is widely used for model prediction in medicine, banking, recommendation systems, threat analysis, and authentication technology. Over time, large amounts of data are collected to provide new solutions to old problems [1]. Large-scale Internet companies collect users' online activities and recommend services of interest to users in the future through the analysis of big data. Health data from different hospitals and government agencies can be used to produce new diagnostic models, while financial companies and payment networks can also combine transaction history, merchant data, and account holder information to train more accurate fraud detection engines [2]. Although the progress of technology at this stage makes the processing and computing of big data more efficient, it is still an important challenge to ensure the privacy and security of cloud data. Competitive advantages, privacy concerns, laws and regulations, and issues surrounding data sovereignty and jurisdiction have hindered the development of data training techniques by many outsourcing

companies [3]. The algorithm workflow of distributed machine learning can be summarized as follows: the system receives large-scale data and stores them in the cloud, and then communicates data in the distributed network. Each distributed computing node performs the corresponding computing task after receiving the required data, and the system aggregates the sub-models trained by each node [4]. The main bottleneck in the work is the privacy security during data training and the model security after each node trains the sub-model, and the efficiency and accuracy of model training cannot be reduced by improving the security. McMahan et al. employed a differential privacy technique on a distributed parallel architecture to enable a trusted server to add noise to the weighted average of user updates to guarantee the user-level privacy [5]. The aggregation scheme of Adadi et al. is proved to be secure in the semi-honest adversary environment, especially when the secure multi-party computation (MPC) computes the sum of individual local user model updates at the cost of computational cost and communication overhead [6]. Shakeel P. et al. proved that when the server is not trusted, differential privacy cannot rely on the server to complete the task of adding noise, and a small part of the original gradient can be used to explain the local data [7]. Li et al. used the federated learning method to protect the user's privacy, but it increased the cost of computation and storage while protecting the privacy security [8]. Elgabli et al. combined the distributed differential privacy with a three-layer encryption protocol and proposed an unbiased coding algorithm to reduce the mean square error to achieve a better trade-off and combination of security and efficiency [9]. This paper is based on the data analysis and calculation of cloud server ciphertext transmission and machine learning distributed training platform. A distributed machine learning scheme (Distributed Machine Learning Privacy-protection Against Collusion Attacks, ACA-DMLP) against collusion attacks is proposed. This comprises the following steps of: adding the Laplace noise disturbance to a ciphertext of a user by a cloud end, and performing disturbance processing on each piece of ciphertext data distributed to a training platform through a differential privacy algorithm. Through the unsolvability of the system of indeterminate equations in the algorithm, the collusion attack of the adversary is prevented. The security evaluation and efficiency performance analysis of the scheme is carried out through simulation experiments. The main innovations of this paper are:

1) This paper proposes a private data encryption scheme that supports multiple users to encrypt private data with

different public keys at any time and upload it to the cloud server to encrypt user data efficiently.

2) Establish a mechanism for the cloud server to add noise to the ciphertext data to efficiently protect the transmitted data.

3) An efficient distributed machine learning scheme is designed, and an anti-collusion attack algorithm is proposed to protect the privacy of each training node, which ensures the security of user privacy data and the training model of each node.

## II. RELATED WORK

### A. Distributed Machine Learning

Distributed Machine Learning is mainly used to study how to use multiple computers to train large-scale data models. Big data has a large volume of data, many types of data, and high commercial value. Big data and cloud computing cannot be separated. With the rise of big data, cloud computing is bound to develop. However, big data cannot be processed by a single computer, so users have to adopt a distributed computing architecture. Therefore, distributed machine learning has also been developed rapidly. However, before the theory and technology related to big data were proposed, there had been a lot of related research work in the industry. In order to make the speed of data calculation and model training faster in machine learning, multiple computers or servers are used to run at the same time. Parallel processing is generally called "parallel computing" or "parallel machine learning". Its main purpose is to decompose a large computing task into multiple small computing tasks, and then distribute them to multiple computers or processing nodes in a distributed architecture for processing and computing. Nowadays, under the dual challenges of large-scale data and large-scale models, there are newer and higher standards and requirements for the computing power and storage capacity of servers used in machine learning:

1) The calculation is more difficult and more complex, so that the previous simple parallel calculation may take a lot of time. Therefore, there is an urgent need for a processor or computer cluster with higher parallelism and computing power to complete the data training task.

2) The volume of data is large and the required storage capacity is large, which leads to the fact that a single machine cannot meet the data storage needs at all, so more and more schemes have to adopt the distributed cluster architecture for data storage.

### B. Differential Privacy Technology

Because of its strong background assumptions, differential privacy has become a mainstream security algorithm in the privacy protection schemes related to machine learning. It can even be said that in the field of cryptography, any algorithm related to privacy protection can use differential privacy [10]. Generally speaking, the most powerful thing about differential privacy is that as long as every step in the algorithm meets the requirements of differential privacy, it can ensure that the final output of the algorithm still meets the requirements of

differential privacy [11].

$m_1$  and  $m_2$  are two adjacent data sets with different records, which are called adjacent data sets (also known as brother data sets). Differential privacy uses the Laplace mechanism to add measurable disturbance to the ciphertext to ensure the security of data distributed by cloud servers [12].

Definition 1:  $\epsilon$ -DP :  $\epsilon$ -DP means that if there is a pair of adjacent data sets  $m_1$  and  $m_2$ , and  $K$  is within the range of  $R$ , then the mechanism  $R$  belongs to  $\epsilon$ -DP, then the following holds:

$$\Pr[R(m_1) = K] \leq e^\epsilon \Pr[R(m_2) = K] \quad (1)$$

Where  $\epsilon$  is the privacy budget, which refers to the number of bits of information that the data analyst DA can obtain. The smaller  $\epsilon$  is, the less bits of information that the data analyst DA can obtain. The stronger the secrecy of  $\epsilon$ -DP is, and the randomness of differential privacy ensures the robustness of differential privacy [13].

Definition 2: Sensitivity:  $f$  is a function in the input space of the data set, i.e.,  $f: m \rightarrow R^d$ , which is used to describe the mapping function of a data set  $m$  to a  $d$ -dimensional space [14].  $\Delta f$  represents the sensitivity of two adjacent data sets, and has the following calculation formula:

$$\Delta f = \max_{m_1, m_2} \|f(m_1) - f(m_2)\| \quad (2)$$

Where, on  $R^d$  with at most one different piece of data, the maximum value is on the pair of  $m_1$  and  $m_2$ .  $\|f(m_1) - f(m_2)\|$  represents the Manhattan distance from a point in the data set in the real domain to a point in the data set  $m_2$ , which is called the 1-norm. For various different pairs of  $m_1$  and  $m_2$  data sets, finding the maximum distance is the sensitivity [15]. Differential privacy means that for a data set with only one record difference, the probability obtained by query is close. The closer the probability is, the stronger the confidentiality of the algorithm to the private data is. If the results of two data queries are completely consistent, the data set has been completely randomized [16]. In this way, the data will lose its availability again and again to improve security. Privacy protection will lose its original role and significance. Most of the schemes make the query probability close, not exactly the same, hoping to find a balance between the security and availability of private data [17].

## III. BIG DATA INFORMATION SYSTEM SCHEME

### A. System Model

The system model diagram of the ACA-DMLP scheme is shown in Fig. 1.



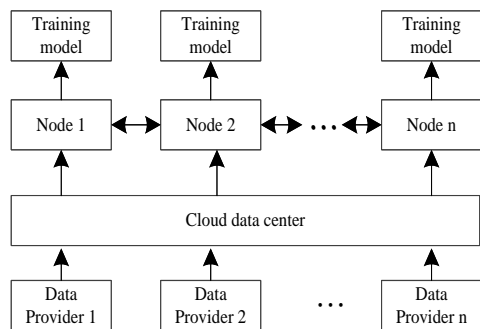


Fig. 1. System Model of the ACA-DMLP Scheme.

A DP is a group of a plurality of different data providers  $DP = \{DP_1, DP_2, \dots, DP_n\}$  in the scheme model, and provides data from a plurality of different sources for a cloud server. The data providers encrypt their own private data and submit the data to a cloud server. The cloud server adds a noise mask to ciphertext data and distributes the ciphertext data according to the requirements of a data analyst. Before the private data set is outsourced to the cloud server for storage, each data provider will use its own public key  $pk_{DP_i} (i = 1, 2, \dots, n)$  to encrypt the sensitive data in its data set, and then entrust it to the cloud server for storage and computation [18].

### B. Scheme Described

While the efficient training is distributed in the working nodes, the privacy data security and the sub-models trained by each distributed node are protected [19]. Fig. 2 is a system flow chart of that ACA-DMLP scheme.

The data provider DP uploads a large number of ciphertext data sets to the cloud server. The cloud server adds noise disturbance to the ciphertext data sets through the differential privacy scheme based on the Laplace mechanism and trains the logistic regression model through multiple iterations of multiple training nodes on the cloud platform. At the same time, it ensures that the adversary colludes with one or more computers in the distributed cluster and will not leak the encrypted data set distributed by the cloud server to the data analyst and the sub-model that has been trained by the computing nodes [20].

To improve the real-time and dynamic performance of data uploaded by users and ensure the security of data, this paper proposes a homomorphic encryption privacy protection scheme, in which each data provider has a public key. The privacy data is dynamically encrypted in real-time by combining the XOR operator of homomorphic encryption and the Diffie-Hellman theory of separable computation. Then upload it to the cloud server. Even if an adversary steals the cloud data, the plaintext cannot be cracked. A hash function is added to the algorithm to ensure the security of the ciphertext. Finally, through the security analysis and proof of the scheme, the feasibility and data security of the scheme are theoretically explained.

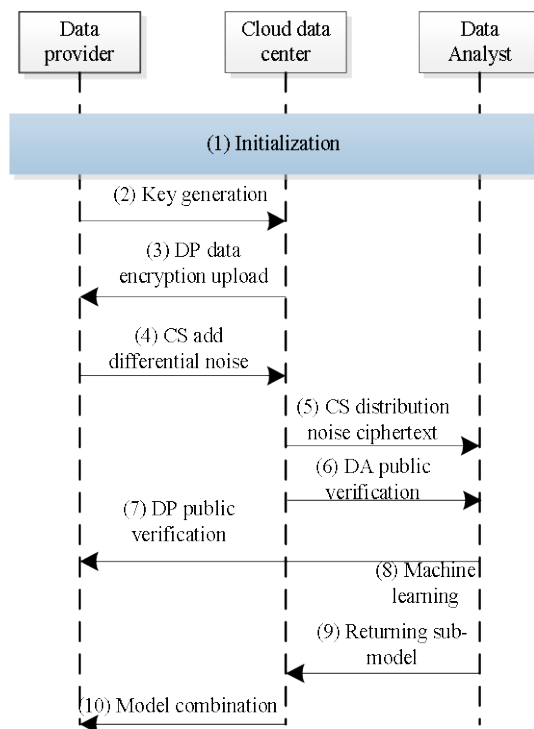


Fig. 2. System Flow Chart of the ACA-DMLP Scheme.

### C. Programme Framework Structure

The main content of the ACA-DMLP scheme is that the cloud server adds noise to the ciphertext data and then distributes the disturbing data to each working node of the data analyst. Due to the unsolvability of the indeterminate equations used by differential privacy to add noise, the adversary cannot theoretically steal the disturbed data through a reverse attack. Because of the particularity of the distributed architecture, the adversary cannot steal the trained sub-model of working nodes by conspiring with one or more hosts [21].

On the coordinate plane, the Manhattan distance between point  $i$  with coordinate  $(x_1, y_1)$  and point  $y$  with coordinate  $(x_2, y_2)$  is calculated as:

$$d(i, j) = |x_1 - x_2| + |y_1 - y_2| \quad (3)$$

If the scheme satisfies differential privacy, if and only if the following expression holds:

$$C(d) = f(d) + Lap\left(\frac{\Delta f}{\epsilon}\right) \quad (4)$$

$C(d)$  is the output function encrypted by the differential privacy algorithm, that is, each data set of the differential privacy algorithm outputs a ciphertext.  $f(d)$  is the ciphertext data received by the cloud server, that is, the cloud server adds noise to the ciphertext to execute the input function  $f(d) = (x_1, x_2, \dots, x_n)^T$  of the differential algorithm. T



represents the transpose of the vector [22]. Converts the input

data set to a vector.  $Lap\left(\frac{\Delta f}{\varepsilon}\right)$  is the noise perturbation added by the cloud server to the encrypted data. It is added in the form of a vector in the scheme. The form of adding noise disturbance is the vector addition operation between the vector of the original ciphertext data set and the noise vector, as shown in Formula 5:

$$C(d) = f(d) + \left( Lap_1\left(\frac{\Delta f}{\varepsilon}\right), Lap_2\left(\frac{\Delta f}{\varepsilon}\right), \dots, Lap_n\left(\frac{\Delta f}{\varepsilon}\right) \right)^T \quad (5)$$

Formula (4) is an algorithm formula for the cloud server to add noise to each piece of ciphertext data (n pieces of data in total) in the scheme. In differential privacy, as usual  $\mu = 0, b = \frac{\Delta f}{\varepsilon}$ , the Laplace function is written as:

$$Lap\left(\frac{\Delta f}{\varepsilon}\right) = \frac{1}{(2\Delta f) / \varepsilon} e^{\frac{-|x|}{(\Delta f) / \varepsilon}} \quad (6)$$

Simplified as follows:

$$Lap\left(\frac{\Delta f}{\varepsilon}\right) = \frac{\varepsilon}{2\Delta f} \exp\left(\frac{-\varepsilon|x|}{\Delta f}\right) \quad (7)$$

The initial ciphertext  $f(d) = (x_1, x_2, \dots, x_n)^T$  is differenced, that is, summed with the Laplace noise vector, to give the following equation:

$$C(d) = (x_1, x_2, \dots, x_n)^T + \left( Lap_1\left(\frac{\Delta f}{\varepsilon}\right), Lap_2\left(\frac{\Delta f}{\varepsilon}\right), \dots, Lap_n\left(\frac{\Delta f}{\varepsilon}\right) \right)^T \quad (8)$$

The following formula can be obtained by substituting the above formula (5) into the vector addition formula (6) and then transposing and expanding it:

$$\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} + \begin{pmatrix} Lap_1\left(\frac{\Delta f}{\varepsilon}\right) \\ Lap_2\left(\frac{\Delta f}{\varepsilon}\right) \\ \vdots \\ Lap_n\left(\frac{\Delta f}{\varepsilon}\right) \end{pmatrix} = \begin{pmatrix} x_1 + \frac{\varepsilon}{2\Delta f_1} \exp\left(\frac{-\varepsilon|x_1|}{\Delta f_1}\right) \\ x_2 + \frac{\varepsilon}{2\Delta f_2} \exp\left(\frac{-\varepsilon|x_2|}{\Delta f_2}\right) \\ \vdots \\ x_n + \frac{\varepsilon}{2\Delta f_n} \exp\left(\frac{-\varepsilon|x_n|}{\Delta f_n}\right) \end{pmatrix} = C(d) \quad (9)$$

Assume that there are N computing nodes in total in the data analyst, and each node allocates i (i is a random number in  $1, 2, \dots, n$ ) pieces of encrypted data. Then the noisy ciphertext data allocated by each working node is:

$$\begin{cases} C_1(d) = f_1(d) + Lap_1\left(\frac{\Delta f_1}{\varepsilon}\right) \\ C_2(d) = f_2(d) + Lap_2\left(\frac{\Delta f_2}{\varepsilon}\right) \\ \vdots \\ C_i(d) = f_i(d) + Lap_i\left(\frac{\Delta f_i}{\varepsilon}\right) \end{cases} \quad (10)$$

$C_N(d)$  is a noise data set allocated by the cloud server to each work node and added with the Laplace noise through differential privacy. Each work node in the data analyst executes related machine learning algorithms such as query, classification, calculation, statistics and the like on the noise data set, and trains a sub-model of each work node with the allocated data. Then the sub-models are submitted to the data center in the cloud server by the working nodes, and the next sub-model is trained, and finally all the sub-models of all the nodes are summarized by the data center to form a complete machine learning model to complete the machine learning task outsourced by the user.

#### IV. SAFETY ANALYSIS

##### A. Data Integrity Analysis

The adversary  $A_r$  attacks the ciphertext and training nodes distributed to the training nodes after the cloud server adds noise. Fig. 3 shows the security model of the ACA-DMLP scheme.

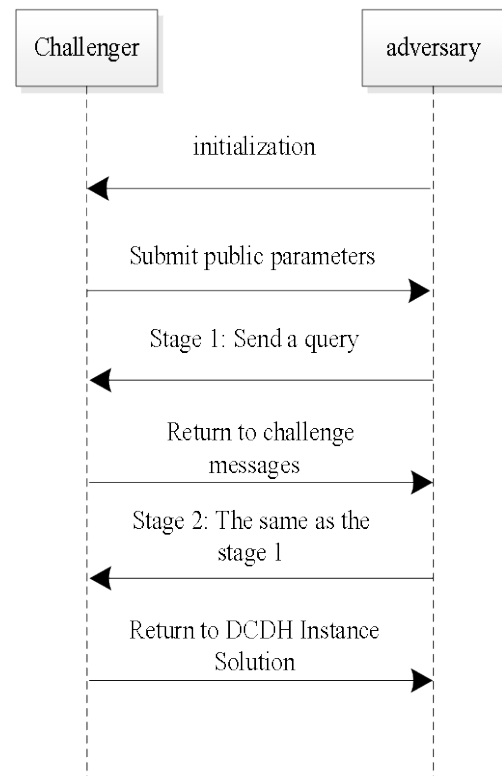


Fig. 3. Security Model of ACA-DMLP Scheme.

Setup:  $B_{tr}$  submits a public parameter  $(q, G, g, H_1, H_2, H_3, H_4, e_0, e_1)$  to  $A_{tr}$ .  $B_{tr}$  uses the list  $(L_{H_1}, L_{H_2}, L_{H_3})$  to simulate the random oracle model of  $H_1, H_2, H_3$  respectively, and guarantees their consistency.  $B_{tr}$  prepares a table  $L_k$  for the public and private keys.

Lemma 2: examine that communication between  $A_{tr}$  and the algorithm  $B_{tr}$  of the scheme in this paper according to the IND-PRE-CCA game.

Phase 1: the adversary  $A_{tr}$  issues a series of queries, and the algorithm  $B_{tr}$  responds to these queries according to the scheme algorithm.

Challenge: The adversary  $A_{tr}$  challenges  $B_{tr}$  to request a ciphertext message  $f(d)$  from the cloud server.  $B_{tr}$  responds to a series of queries from  $A_{tr}$  with a ciphertext  $C_N(d)$  that contains  $x_n$  and  $\frac{\varepsilon}{2\Delta f_n} \exp(\frac{-\varepsilon|x_n|}{\Delta f_n})$ . Due to the unsolvability of the indeterminate system of equations, the adversary cannot infer the initial ciphertext  $f(d)$  from  $C_N(d)$ .

Phase 2:  $A_{tr}$  continues to issue attack queries as in Phase 1, and algorithm  $B_{tr}$  continues to respond to adversary L's queries in the challenging manner described above.

Guess:  $B_{tr}$  returns a solution to the DCDH instance. In the random oracle model, the scheme is secure under the IND-PRE-CCA property. If an adversary  $A_{tr}$  corrupts CS or DA to obtain the outsourced data,  $A_{tr}$  cannot get the plaintext due to the IND-PRE-CCA nature of the scheme. In addition, if  $A_{tr}$  gains access to some data, the scheme achieves  $\varepsilon - DP$  due to the Laplace mechanism adding noise and the unsolvability of the algorithm equations. Therefore, the scheme is secure under the random  $\varepsilon - DP$  model.

### B. Collusion-Resistant Analysis

Due to the semi-honesty of the training nodes in the data analyst, suppose that there are  $\delta(1 \leq \delta \leq N)$  training nodes in the data analyst who collude with their training submodel  $R_i$  to steal  $R_e = (R_1, R_2, \dots, R_\delta)$ . Due to the strong background assumption of differential privacy, at least one training node does not participate in the collusion attack, and the adversary solves the logarithmic equation  $C(d) = f(d) + Lap(\Delta f / \varepsilon)$ . Because of its difficulty, the adversary cannot solve  $1 \leq i \leq n$  and  $\Delta f_i$ . Assume that the

adversary guesses  $\Delta f_i$  after several repetitions with a very low probability. According to equation (7), the adversary conspires to construct a system of equations with  $\delta$  equations and  $\delta + 1$  unknowns:

$$\begin{cases} C_1(d) = x_1 + \frac{\varepsilon}{2\Delta f_1} \exp(\frac{-\varepsilon|x_1|}{\Delta f_1}) \\ C_2(d) = x_2 + \frac{\varepsilon}{2\Delta f_2} \exp(\frac{-\varepsilon|x_2|}{\Delta f_2}) \\ \vdots \\ C_\delta(d) = x_\delta + \frac{\varepsilon}{2\Delta f_\delta} \exp(\frac{-\varepsilon|x_\delta|}{\Delta f_\delta}) \end{cases} \quad (11)$$

Since Equation (9) is an indeterminate equation system with infinitely many solutions, the adversary cannot solve all equation unknowns through the limited unknowns. The adversary cannot conspire to calculate all initial ciphertexts  $f(d)$  and all sub-models  $R_i$  through Equation (7).

## V. EXPERIMENTAL RESULTS AND ANALYSIS

### A. Experimental Setup

In the scheme, the data set is stored in the cloud server, so the user's local storage cost is small. The main analysis is not done in the scheme. The cost of the scheme depends on the time overhead, including encoding time, communication time, and computation time. To evaluate the time cost of the scheme in this paper, the test platform is shown in Table I.

The experiment examines the running time of the relevant scheme on the MNIST data set of size  $(m, d) = (12396, 1568)$ , where m is the sample size of the training data set and d is the test sample size. The distributed computing is simulated by the platform without considering the network delay. The time cost and accuracy of the proposed scheme are compared with the schemes in [23], [24] and [25].

### B. Efficiency Analysis

In the experiment, the total time cost of the scheme in this paper and the schemes in [23], [24] and [25] is simulated and analyzed by setting the number of training nodes  $N = (5, 10, 15, 20, 25, 30, 35, 40)$ . Fig. 4 compares the data calculation efficiency of the four schemes under different numbers of training nodes.

TABLE I. TEST PLATFORM CONFIGURATION

Type	Settings
CPU	Intel(R) Core(TM) i7-10700 4.8GHZ
RAM	32G DDR4
Hard disk	1T SSD
Operating system	Windows10
Data set	MNIST
Simulation platform	Python

In the experiment, the training time of different schemes is measured while the number of nodes is gradually increased. The following conclusions can be drawn: when  $N = 5$ , due to the small number of nodes and the small degree of parallelism between hosts, all schemes have almost the same performance; With the increase in the number of nodes, the difference in the time spent by different schemes to process the same amount of data sets is increasing, and the performance comparison between schemes is more obvious. Due to the distributed nature of the proposed scheme, the total amount of data sets are evenly distributed in each node. The more the number of training nodes is, the less time it takes to process the same task, and the shorter the total running time is. Thus, the time to process data sets decreases with the increase of the number of nodes. Compared with the scheme in the reference, the time cost of the scheme in this paper is smaller. Therefore, through the comparison of Fig. 4, it can be seen intuitively that the scheme adopted in this paper has obvious computational advantages. In the MPC scheme of [25], no matter how many hosts there are, each host repeatedly computes the entire data set to meet the needs of processing all tasks, so the computing time tends to increase. Through analysis, when the number of hosts is  $N = 40$ , it is found that the scheme in this paper has a significant improvement in efficiency compared with the schemes in [23], [24] and [25]. Fig. 5 shows the comparison of communication time (Comm), encoding time (Enco), computation time (Comp) and total time (Total) between the proposed scheme and the reference scheme when  $N = 40$ .

It can be seen from the images that the running time of each part of this scheme has been significantly improved compared with the reference schemes [23], [24] and [25]. The main reason is that the user encrypts the private data through the homomorphism in the scheme, which simplifies the algorithm and reduces the complexity. In the reference scheme [25], the data set size of each host is the same as the original data set, while the data set of each host in the present scheme is only 1/40 of the original data set. This is because the distributed machine learning provides a large parallelization gain for the scheme, while the reference scheme has a large computational overhead.

### C. Accuracy Analysis

MNIST data set is set in the experiment (12396 samples are used in the training set, and 1568 samples are used in the test set). Since [24] does not discuss the problem of training accuracy, the accuracy of this scheme is compared with that of the schemes in [23] and [25]. When the number of hosts is  $N = 40$ , the accuracy of this scheme is compared with that of the schemes in [23] and [25] under different iterations. Fig. 6 illustrates the experimental comparison results of the three schemes.

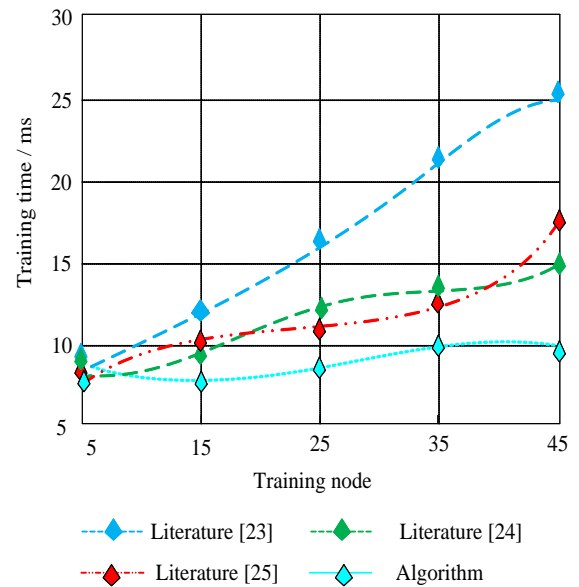


Fig. 4. Comparison of Running Time of Different Schemes.

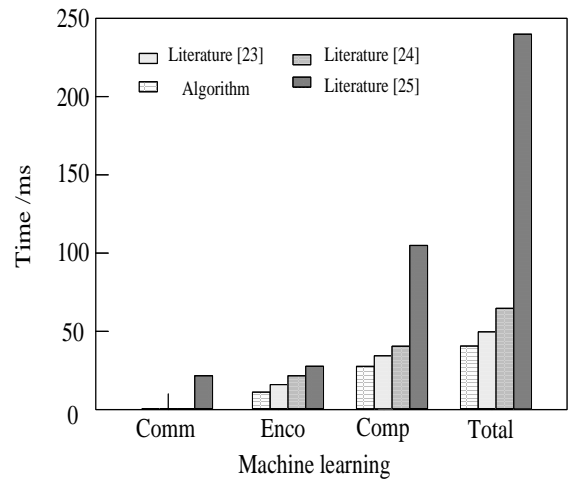


Fig. 5. Time Comparison of each Link of Different Schemes.

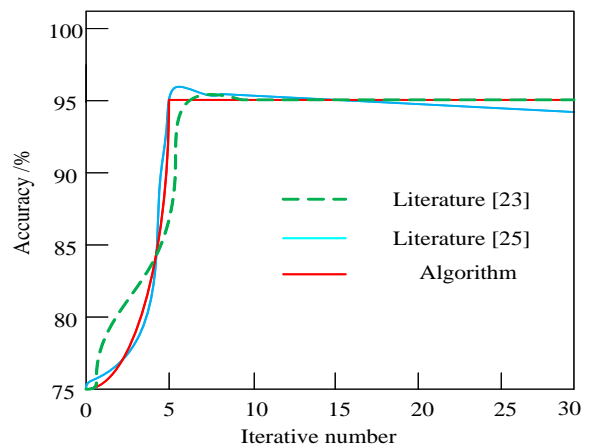


Fig. 6. Comparison of Accuracy of the Three Schemes.

It can be seen from Fig. 6 that during model training, compared with the reference schemes [23] and [25], this scheme has the same number of iterations, and the accuracy of model training of the two schemes is the same. When the number of iterations is five, there is a slight difference in the training accuracy of the scheme in this paper. The training accuracy of the scheme in this paper is slightly different from that of the comparison scheme, but the difference is kept within 2%. With the increase in the number of iterations, the accuracy between the schemes is getting closer and closer. The difference between the training accuracy of the scheme in this paper and that of the scheme in the reference is only 0.2%. This result shows that the scheme used in this paper almost guarantees the same accuracy as the reference scheme when the data set is unchanged and the number of iterations of the three schemes is the same. This scheme does not reduce the accuracy of model training because of the improvement of computational efficiency and data security.

#### D. Discussion

By comparing the functions of each scheme, it can be seen that [23] uses distributed architecture to improve the efficiency of data training, and uses homomorphic encryption algorithm and differential privacy to ensure the security of user privacy, and supports the public verification of ciphertext by each entity in the model. Appropriate measures are not taken to defend against the collusion between the adversary and the training nodes. The author in [24] uses a distributed model to speed up data analysis and improve the efficiency of training, but does not take security algorithms to protect the security of user data. The scheme in [25] can resist the collusion attack of the adversary and the training nodes in the distributed scheme, but it does not support ciphertext operation and public verification, and does not use differential privacy technology to protect the security of user privacy data. This scheme uses distributed structure to shorten the time of data analysis and improve the efficiency of machine learning, and uses homomorphic encryption algorithm to support the training platform to train on the ciphertext, uses differential privacy to strictly prevent the user's private data from being leaked in the process of transmission and training, and prevents the collusion theft of adversaries and distributed training nodes. At the same time, each role in the model is supported to download and publicly verify the data in the ciphertext domain at any time to ensure the integrity of user privacy data. Through the above analysis and comparison, the scheme in this paper has high feasibility, and strictly guarantees the integrity of user data, and improves the training efficiency of machine learning.

#### VI. CONCLUSION

In this paper, a collusion-resistant distributed machine learning privacy-preserving (ACA-DMLP) scheme is proposed.

1) The scheme adopts the architecture of distributed machine learning and improves the efficiency of data training through the cluster parallel systems.

2) A differential privacy encryption algorithm and a Laplace mechanism are used to add noise disturbance to the

ciphertext data in the cloud server to ensure data security in the ciphertext domain.

3) The feasibility and high efficiency of the scheme are objectively proved by simulation experiments on relevant platforms. The scheme in this paper improves the security and analysis efficiency of user private data in machine learning and can prevent adversaries from colluding with semi-honest working nodes within data analysts to steal data. The scheme in this paper only considers the safety of the model before training and the sub-model of the training node in machine learning but does not make an effective scheme analysis and demonstration of the data processing, model combination, and the safety of the overall model after machine learning. In future work, the data after training will be processed safely and the outsourcing agencies will be guaranteed to submit the machine learning results to users safely.

#### CONFLICT OF INTEREST

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

#### DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available upon reasonable request.

#### REFERENCES

- [1] L. Cao, Y. Kang, and Q. Wu, "Searchable encryption cloud storage with dynamic data update to support efficient policy hiding," *China Communications*, 2020, 17(6): 153-163.
- [2] K. Liu, J. Peng, and J. Wang, "Scalable and adaptive data Replica placement for geo-distributed cloud storages," *IEEE Transactions on Parallel and Distributed Systems*, 2020, 31(99): 1575-1587.
- [3] Samankumara, Hettige, and Eshani, "Usage of cloud storage facilities by medical students in a low-middle income country, Sri Lanka: a cross sectional study," *BMC Medical Informatics and Decision Making*, 2020, 20(1): 1-8.
- [4] S. Jing, A. Ebadi, and D. Mavaluru, "A method for virtual machine migration in cloud computing using a collective behavior-based metaheuristics algorithm," *Concurrency and Computation: Practice and Experience*, 2020, 32(2): 1-13.
- [5] I. Filip, A. Potoac, and R. Stochioiu, "Data capsule: representation of heterogeneous data in cloud-edge computing," *IEEE Access*, 2019, 7: 49558-49567.
- [6] I. Mavridis, and H. Karatza, "Combining containers and virtual machines to enhance isolation and extend functionality on cloud computing," *Future Generation Computer Systems*, 2019, 94(5): 674-696.
- [7] P. Shakeel, S. Baskar, and H. Fouad, "Internet of things forensic data analysis using machine learning to identify roots of data scavenging," *Future Generation Computer Systems*, 2021, 115: 756-768.
- [8] A. Jeavons, "What is artificial intelligence," *Research World*, 2017, 2017(65): 75-75.
- [9] E. Tapoglou, E. Varouchakis, and I. Trichakis, "Hydraulic head uncertainty estimations of a complex artificial intelligence model using multiple methodologies," *Journal of Hydroinformatics*, 2020, 22(1): 205-218.5.
- [10] K. Ishii, "Comparative legal study on privacy and personal data protection for robots equipped with artificial intelligence: looking at functional and technological aspects," *AI & society*, 2019, 34(3): 509-533.

- [11] K. Lin, J. Lu, and C. Chen, "Unsupervised deep Learning of compact binary descriptors," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019: 1-1.
- [12] W. Li, B. Jiang, and W. Zhao, "Obstetric imaging diagnostic platform based on cloud computing technology under the background of smart medical big data and deep learning," *IEEE Access*, 2020, (99): 1-1.
- [13] W. Liu, J. Guo, and F. Yao, "Adaptive protocol generation for group collaborative in smart medical waste transportation," *Future Generation Computer Systems*, 2020, (110): 167-180.2.
- [14] U. Boryczka, and M. Bachanowski. "Using differential evolution in order to create a personalized list of recommended items," *Procedia Computer Science*, 2020, (176): 1940-1949.
- [15] Helmi, Abrougui, Habib, "Autopilot design for an autonomous sailboat based on sliding mode control," *Automatic Control and Computer Sciences*, 2019, 53(5): 393-407.
- [16] D. Pal, and C. Arpnikanondt, "Analyzing the adoption and diffusion of voice -enabled smart-home systems: empirical evidence from Thailand," *Universal Access in the Information Society*, 2020: 1-19.
- [17] U. Udhayakumar, and G. Murugaboopathi, "To improve user key security and cloud user region-based resource scheduler using rail fence region-based load balancing algorithm," *Journal of Ambient Intelligence and Humanized Computing*, 2020(6): 1-8.
- [18] P. Puteaux, M. Vialle, W. Puech, "Homomorphic encryption-based LSB substitution for high-capacity data hiding in the encrypted domain," *IEEE Access*, 2020, 8: 108655-108663.
- [19] D. Alistarh, D. Grubic, and J. Li, "Qsgd: Communication-efficient sgd via gradient quantization and encoding," *Advances in Neural Information Processing Systems* 30, Curran Associates, 2017, (2017): 1709–1720.
- [20] C. Gao, Q. Cheng, and X. Li, "Cloud-assisted privacy-preserving profile-matching scheme under multiple keys in mobile social network," *Cluster Computing*, 2018, (2018): 1655-1663.
- [21] P. Li, T. Li, and Y. Heng, "Privacy-preserving machine learning with multiple data providers," *Future Generation Computer Systems*. 2018, (87): 341-350.
- [22] Z. Wei, J. Li, and X. Wang, "A lightweight privacy-preserving protocol for VANETs based on secure outsourcing computing," *IEEE Access*, 2019, 7 (99): 62785-62793.
- [23] H. Alzubair, and H. Rafik, "An efficient outsourced privacy preserving machine learning scheme with public verifiability," *IEEE Access*. 2019, (7): 146322-146330.
- [24] C. Fang, Y. Guo, and N. Wang, "Highly efficient federated learning with strong privacy preservation in cloud computing," *Computers & Security*, 2020, 96: 101889.
- [25] Y. Hana, L. Michael, A. Said, "Systematic review on fully homomorphic encryption scheme and its application," *Recent Advances in Intelligent Systems and Smart Applications*, 2018, (2020): 537-551.

# A Reusable Product Line Asset in Smart Mobile Application: A Systematic Literature Review

Nan Pepin<sup>1</sup>, Abdul S. Shibghatullah<sup>2\*</sup>, Kasthuri Subaramaniam<sup>3</sup>, Rabatul Aduni Sulaiman<sup>4</sup>, Zuraida A. Abas<sup>5</sup>, Samer Sarsam<sup>6</sup>

Institute of Computer Science & Digital Innovation, UCSI University, Cheras, Kuala Lumpur, Malaysia<sup>1,2,3</sup>

Department of Software Engineering, Faculty of Computer Science & Information Technology, Universiti Tun Hussein Onn Malaysia, Batu Pahat, Johor, Malaysia<sup>4</sup>

Department of Intelligent Computing & Analytics, Faculty of ICT, Universiti Teknikal Malaysia Melaka, Melaka, Malaysia<sup>5</sup>  
School of Strategy and Leadership, Coventry University, United Kingdom<sup>6</sup>

**Abstract**—A reusable product line asset is a product or asset that can be reused for different purposes including charity. Smart mobile applications are one of several communication and information methods used in charitable activities. Web, mobile, or hybrid platforms can be used to develop charity applications. It takes design and purpose to build an application, whether methodology or software development is applied for the smooth design or development of an application. The data for this study were acquired from the appropriate literature between 2017 and 2021 in order to determine the application development on current charity applications. The Systematic Literature Review (SLR) was employed in this study. The SLR method is used to identify, review, evaluate, and analyze all available research on relevant topics, as well as research issues for philanthropic development. This study aims to answer the following research questions: identify the donation applications that are frequently developed by researchers; identify the methods that are commonly used in the development of charity applications; identify the application platform that is frequently used; identify the functions utilized to the developed application and identify the key users who are using the application. The findings show, charity donations app, structured method, mobile applications, authentication and charity centers and donors were the most often observed in this study.

**Keywords**—Application; charity; donation; reusable product line; systematic literature review

## I. INTRODUCTION

The advancement of information technology is very rapid nowadays. The growth of the telephone or mobile phone, as well as the internet's existence, has had an impact on many facets of life. Whether in one's personal life, socially, or in relation to the world of business or business. Information technology is used in philanthropic activities in addition to facilitating and speeding up communication and information processes. Humanitarian and social concerns can be carried out more quickly by taking advantage of technical advancements. This is because the internet, which has become a part of daily life, makes it easier to access information [1] [2]. The society has a role to play in the expansion of universal social and humanitarian concern activities. As a result, charitable activities have the ability to be done conveniently and quickly through the use of internet technology [3]. Many websites and mobile charity

applications have sprung up to function as a middleman, connecting people who want to help such campaign organisers and donors with people who need support [4]. With the growth of charity platforms nowadays makes it easier for everyone who wishes to donate or seeks donations with only access to the internet. Many communication and information technologies, like an application, are being created to assist users in undertaking donation activities. Application development is often classified into three platforms: desktop, online, and mobile. It takes a good planning to build an application, which must be examined to see whether the purpose, software development process used for the smooth design or development of an application. As a result, the objective of this study is to discover and examine prior studies relevant to the development of donation applications developed by previous researchers. The findings from this study will be used in the following phases for the proposed application SeekandHelp. Various stages should be evaluated before designing an application, including the area of focus, purpose, system user, method, or platform on which the application will run. The systematic literature review (SLR) strategy was applied to discover relevant findings and identify publications in a systematic way, with each procedure according to predefined phases or criteria [5]. This study is important to determine the application development on current charity applications so that it will give researchers to gain some insights on the application. The following sections in this study are organized as follows. Section II describes the methodology used for comparison study, Section III presents the result and analysis and lastly the conclusion.

## II. METHODOLOGY

This section will discuss and describe the phases of research process through using SLR method. SLR is a process of identifying, evaluating, and interpreting all research sources relevant to a research question regarding a research topic. Primary studies and secondary studies are studies and publications that contribute to SLR.

### A. Systematic Review Process

The systematic literature review method conducted will follow a specific process [6]. Each stage of this process will perform its respective tasks according to the order of research to obtain systematic research results. The process begins with

identifying research problems followed by formulating research questions and objectives. The next stage is the literature search strategy stage which process is to determine what keywords to use in article searching in the database and what searching criteria to include and exclude. In next stage, after search strategy completed is the process selection of the study which the process will including identifying, selecting, literature, extracting and evaluating the reporting research results obtained with followed by the last stage in this process. The systematic review process can be seen as Fig. 1.

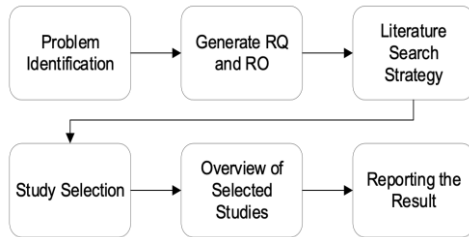


Fig. 1. Systematic Review Process

### B. Problem Identification

The initial stage for systematic review is to identification of the research problems. Research problems are issues raised in related research which will be examined to be discussed and resolved. In this study, the research problem identified is that there is still a lack of research in the current study that provides a structured and comprehensive literature review regarding the development of applications in charity causes.

### C. Generate RQ and RO

In this stage, Research Questions (RQ) and Research Objectives (RO) are created. Research questions function to extract detailed information from each study conducted. Four research question formulations were made related to the study's namely the extent to which the implementation and development of applications related to charity were developed. The four questions have been generated and research questions are arranged in Table I.

TABLE I. RESEARCH QUESTIONS (RQ) AND RESEARCH OBJECTIVES (RO)

Research Question (RQ)	Research Objective (RO)
RQ1. What are the purpose and area of charity researcher focused in developing charity app?	Identify the purpose and types of topics focused on by researchers in developing the applications for charity donation.
RQ2. What methods were used in the app development of the study?	Identify method used by researchers. Knowing the method utilized in the various studies, this information can serve as a point of reference for designing a donation application for use in further study.
RQ3. What platforms are widely used in the development of the charity application?	Identify the platforms most frequently used to develop charity apps. Therefore, it is necessary to consider what platforms can be used for further studies.
RQ4. What functions and features are suggested in the charity app developed?	Identify the features used to develop charity apps
RQ5. Who is the focused user in system development in these studies?	To develop a system, it is necessary to identify who is the user will be involved.

### D. Literature Search Strategy

At this stage, the search process is carried out to obtain relevant sources to answer the Research Question (RQ) and other related references. The search process is carried out using electronic search journal databases, namely, Google Scholar and IEEE Explorer. This search on Google Scholar was chosen because Google Scholar provides a simple way to broadly search for scholarly literature and quickly gauge the visibility and influence of recent articles in scholarly publications

The keyword "local charity donation application" was used as the search query/search string. The search query or search string is what's utilized to find published material that explains the process of creating applications for charity causes. The keywords are formalized by using Boolean operators, namely the process of finding information from queries using Boolean expressions [7]. The Boolean expressions use the logical operators AND, OR and NOT in determining the calculation results only in the form of binary values. The Boolean retrieval result is only relevant documents or nothing. Thus, a Boolean retrieval overload does not result in the same document. Fig. 2 illustrates the search string.

1) *Literature inclusion and exclusion criteria:* Inclusion and Exclusion Criteria is performed to decide whether the data found is suitable for use in SLR research or not. Article search is selected based on criteria such as year of publication, type of journal article, books and proceedings. The Inclusion and Exclusion criteria's in this study is show as follows Table II.

### E. Study Selection Process

The procedure by which records identified in the preceding stage are evaluated for inclusion in the study is described as study selection. By objectively and methodically applying the predetermined eligibility criteria to each record to evaluate if the article should be included, selection bias can be avoided. This study selection process is shown in Fig. 3.

(Localized OR local )  
AND  
(Charity OR Charitable OR Donation)  
AND  
(Application OR System)

Fig. 2. The Finalized Search String/Sentence.

TABLE II. INCLUSION AND EXCLUSION CRITERIA

Inclusion Criteria	Exclusion Criteria
Literature publication from 2017 to 2021.	Literature publication year is below 2017.
Literature focusing on application development with related topics.	Literature which not discusses related topics in the study will not be selected.
Literature that uses only English.	Literature is not in English.



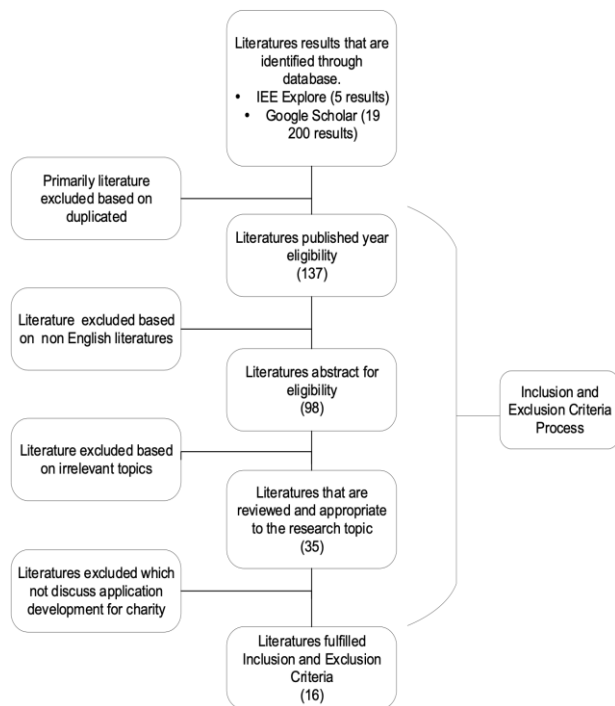


Fig. 3. Study Selection Process.

The selection process begins after obtaining literature search results in the database and obtained articles with details from Google Scholar with results with total of 19200 articles and in IEEE Explorer resulted in five literatures. Because of the large number of results, the first 250 literatures are selected for the screening stage. The first primary and second screening process is to exclude literatures based on duplicate literatures. The initial and second screening generated results of 137 and 98 literatures are excluded from non-English and year eligibility. The next screening is on study and understanding the abstracts of the literature to identify which articles are appropriate and which are not in accordance with the criteria for inclusion. Based on a comprehensive review of the title and abstract, there are 35 study literatures that meet the criteria relevant to the topic discussion. The following screening assessing of each literature for fully inclusion and exclusion criteria examines the literature in more depth. The process's inclusion and exclusion criteria finally resulted of 16 literatures accessible for this study.

1) *Data extraction and analysis:* The selected article then performed was data extraction to answer the research question (RQ). The data extracted from the research article discusses the development of charitable applications, whether the purpose, area, development methods, platforms used, features of the purpose app and system users of each application developed. By reading and analyzed the selected articles, an analysis will be produced that can answer the research questions. This process will be described in the results section and subsequent discussion in detail.

### F. Reporting the Result

Select Studies Overview indicates the distribution of selected literature through that source published. From 1320 literatures, 34 literatures were selected as shown in Table III.

## III. RESULT AND DISCUSSION

RQ1. What are the purpose and area of charity researcher focused in developing charity app?

The applications were grouped by the area focused by research in the literatures show in the Table IV divided into four groups: (1) Charity center donation, (2) Blood donation finder apps, (3) Food donation app, (4) Crowdfunding donation apps, and (5) Donation handling apps.

### A. Charity Center Donation

A total of six literatures were focused develop an app to find charity center for donation. The apps share the information of charity center to receive of donations and with information of center will be easier to find donors to meet the needs of the charity center. As studies by [13] focused on application to help the donor to find information about the non-profit organization available, [10] develop applications to facilitate donors and organizations to reach out directly and [8] reduce poverty. Study on [12] was not just to facilitate social institutions in making donations, but in [9] the apps were helpful to connect people to donate their used items and meanwhile, [11] help to reduce wastage and fulfil other items requirements of needy organizations. In the studies, an application was developed for further maximizing social services in the development of social welfare. The author of the literature focused on donation application design to facilitate the community in donating to the charity center in various forms, such as used good, volunteer, etc.

TABLE III. AREAS OF CHARITY DONATION APP FOCUSED

Focused donation area's	Ref. No
Charity donation center app	[8], [9], [10], [11], [12], [13]
Donation handling process app	[14], [15]
Blood donation app	[16], [17], [18], [19], [20]
Crowdfunding donation app	[21], [22]
food donation app	[23]

TABLE IV. DESCRIPTION OF SELECT STUDIES OVERVIEW

Literature	Purpose	Method	Platform	Area of Donation system focused	Features	User
[8]	To facilitate people and organizations to reach out easily in Bhutan.	Structured method	Web Application	Charity donation center app	Authentication	Charity organization and donor
[14]	To provide a better solution for donation handling in Sri Lanka during disasters.	Structured method	Hybrid	Donation handling process app	Authentication and Geolocation	Charity organization, donor and admin
[16]	To fill the knowledge gap about new technologies that can be developed to help the community in relation to blood finding.	Structured method	Mobile Application	Blood donation app	Geolocation	Donor and community
[9]	To facilitate and connect people in books donation.	Structured method	Mobile Application	Charity donation center app		Donor and seeker
[10]	To help people donate items in an easy way to the charitable organization in Malaysia.	Structured method	Mobile Application	Charity donation center app	Authentication	Donor and charitable organization
[21]	To design the crowdfunding platform specifically tailored to address food. insecurity problems in Indonesia.	Structured method	Mobile application	Crowdfunding donation app	Authentication	Charity organization, community
[20]	To improve the current existing system regarding the blood donation.	Structured method	Hybrid	Blood donation app	Authentication	Donor, patient, admin and authorize organization
[11]	To provides a platform for donation of useful items to the nearest NGO and reduce the wastage.	Object oriented	Mobile Application	Charity donation center app	Authentication and Geolocation	Charitable organization and donors
[12]	To help charitable organization and donor to raising/ giving donations.	Structured method	Mobile Application	Charity donation center app	Geolocation, Authentication and notification	Charitable organization and donors
[23]	To design and facilitate regarding food security issues.	Food donation	Mobile Application	food donation app	Authentication	Donor and community
[15]	To build an application program to reduce manual work of managing for the blood donation process	Object oriented	Mobile Application	Donation handling process app	Authentication and Geolocation	Charity organization, admin and donor
[17]	To develop a mobile blood donation management system and enhance the existing system.	Object oriented	Mobile Application	Blood donation app	Authentication and Geolocation	Authorize organization and donor
[18]	To solve the problem of searching for blood donors in the city of Lampung.	Object oriented	Mobile Application	Blood donation app	Geolocation	Donor, community and admin
[22]	To design, develop and test the decentralized crowdfunding web application in facilitating the donation process on Ethereum network.	Object Oriented method	Web Application	Crowdfunding donation app	Authentication	Charity organization and donor
[13]	to solve the problem by providing the donor complete information about the organization that accept the used item for donation.	Structured method	Mobile Application	Charity donation center app	Notification Authentication	Charity organization and donor
[19]	To develop blood donation app for handle emergency situation of blood availability.	Object Oriented method	Mobile application	Blood donation app	Authentication	Donor, Admin and authorize organization

### B. Blood Donation App

Research focusing on the development of applications for blood donation was obtained from relevant studies. This app focused on as an information tool to find blood donors or donate their blood. The blood application serves to provide information to users in facilitating the monitoring of blood supply and coordination of the processes involved. The availability of appropriate bloodstock is crucial for use during special medical condition [25]. Thus, various apps are design to solve the problems. Six articles were identified which researchers [16] and [18] develop an app to focus on blood donors to aid the community in relation to blood donation finders. A study by [20] focused to improve the existing system which was studied facilitating monitoring of blood supply and coordination of the processes involved. The research on [18] focused on to solve the problem of searching for blood donors based on the closest location. In [17], to the factor of authenticity, the apps develop focus to facilitate the work of the process of managing donors' information more quickly while [19] with organized blood bank donation data and information on bloodstock will facilitate people with easier access to bloodstock information without visiting a blood bank.

### C. Food Donation App

Researcher focused on food donation app which is a social movement themed application of donating food waste and foodstuffs to urban communities, in an effort to develop a digital community in reducing the amount of food waste. Research by [23] developed food donation applications as intermediaries to address food insecurity and eliminate food waste by using food sources available in local communities. This food -focused research is how to design an app with the theme of a social movement donating food and food waste to the community, in an effort to develop a digital community in reducing the amount of food waste.

### D. Crowdfunding Donation App

Crowdfunding is a fundraising practice for various types of businesses, whether in the form of product ideas, businesses or profits derived from donations from the general public or groups. There are two studies focusing on crowdfunding donations which researcher in [22] developed a decentralized crowdfunding focused application to facilitate the donation process on the network and record donation transactions securely. While researcher in [21] developed a platform specifically designed to address food insecurity issues that allow the public to make contributions to achieve targeted campaign funding. This type of development is different from regular donations, donations in crowdfunding are done online, where digital money is given for each campaign or project undertaken.

### E. Donation Handling Process App

Two studies found which focusing on developing an app for the process handling of donations, namely, researcher [14] which develops applications to improving the process of handling donations when a disaster strikes. It provides a better solution for handling donations when they are needed by the community. Meanwhile, researcher in [15] developed an application that focuses on the process of reducing the manual

work of managing donations in charity organizations in helping social organizations to run projects more transparently.

RQ2. What method were adopt in the development of application in the study?

Table V show structured methods which are the dominant method in the development of charitable applications. Structured methods in information system development, often known as the System Development Life Cycle (SDLC). In general, SDLC is sequentially divided into six stages, namely, planning, analysis, design, implementation and maintenance. This structured method is made up of many models, and in this study, we found as total 10 researchers who utilize it in the development of charitable applications.

An agile model is used in studies [13], in which the work process is repeated in an organized and systematic way. The prototype model has been utilised in studies [8], [14], [9], [20] and [23] which this model allows to understand the requirements at an early stage of development, helps get valuable feedback to understand what exactly to expect from the product under development. The authors in [16] and [21] utilise spiral model, which supports risk management and this model consists of multiple cycles, similar to a combination of waterfalls and iterations divided into each phase. The Rapid Application Development (RAD) paradigm is utilised in this study by [10] to emphasise rapid prototyping and quick feedback within an ongoing development and testing cycle, and it may produce multiple iterations and update software quickly with fast application development. In [12], the author utilised the waterfall model or known as classical method that follows a gradual regular pattern worked from top to bottom.

Object-oriented methods, as opposed to structured methods, total six researchers were utilized in their studies [11], [15], [17], [18], [19] and [22]. This object oriented method is used to simplify all kinds of problems that exist in a system by using many objects.

RQ3. What platforms are widely used in development of application in the studies?

The results shown in Table VI show that the dominant platform used in developing charity-based applications is the Mobile Application. Research focused on mobile platforms was carried out by [10], [16], [9], [21], [11], [12], [15], [17], [18], [13], [19], [15], [17], [18], [13] and [19].

TABLE V. METHOD USED IN SYSTEM DEVELOPMENT

Method	No. Ref
Structured method	[8], [14], [9], [20], [23], [16], [21], [12], [10] and [13]
Object Oriented method	[11], [15], [17], [18], [19] and [22]

TABLE VI. PLATFORM USED TO DEVELOP AN APPLICATION

Platform	Ref. No
Mobile Application	[10],[16], [9], [21], [11], [12], [15], [17], [18], [13], [19] and [23]
Web application	[8], [22]
Hybrid (Web and Mobile)	[14], [20]

A mobile app is an application that can be put into a mobile device and used at any time and from any place [21]. This is because mobile applications and their global influence show that they are used and impacted differently by individuals, businesses, and social groups [16]. Users can take advantage of the ability to use and benefit from mobile devices that can be accessible by mobile devices over mobile telecommunications networks [12], which helps users get to their desired place in their preferred time slot [11] and provides additional e-philanthropy experiences [10]. As a result of its simplicity of use, ubiquitous availability, and speed, smartphones have become more of a necessity than a commodity for most people [9][24]. Research on [8] and [22] applying web application development to the charities studied. Web application platforms, according to researchers, have lower communication costs. It can also encourage people to participate more often as there are no time constraints [22]. Meanwhile research by [14] and [20] focuses on both mobile and web application development. Implement a hybrid system makes it easier for users and organizations to gain access to the system [14]. Aside from that, mobile application and a website hosted on a cloud hosting that serves as an interface for system users and leverages the cloud to store and process data. This enables the system to take use of cloud hosting characteristics including rapid deployment, high availability, scalability, and management simplicity [20].

RQ4. What are the functionalities and features proposed in the development of charity apps?

This question is discussed by considering the results obtained from the implementation of 16 applications selected and answered the questions of RQ4 mentioned above. The features were evaluated for each of the selected literature. The results are presented in the Table VII.

#### F. Authentication

Authentication is the process of gaining recognition or gaining recognition. So this validation will verify who has actually interacted with the system. Authentication implementations can include what only the person knows [8], [10], [21], [12], [11], [15], [23], [17], [17], [19], a user who logs in with a username and password that only he knows, can be ascertained that that person is actually logged into the system when it comes to biometric authentication which involves part of body to be recognized by the user [26]. Researcher in [14] on the use of biometrics, for example, facial identification using the camera used to identify and record the faces of the people in an array for future accesses or [20] to identify and measure the health.

TABLE VII. COMMON FEATURES PROPOSED

Feature//Function	Ref. No
Authentication	[8], [10], [21], [12], [11], [15], [23], [17], [17], [19], [14]
Notifications	[12], [13]
Geolocation	[14], [16], [15], [17], [18], [12], [11]

#### G. Notifications

There are two applications that have been developed to provide users with a notification feature [12], [13]. This feature's implementation allows users to be notified in a variety of ways. Among those investigated, a notification could be an immediate alert the user receives about a reminder, news, or the next date on which event is possible. Users can also get push notifications for upcoming campaigns in their area as well as instant invitations to charitable events. Notifications are a potent tool that takes attention away the user's attention. However, if these notifications convey irrelevant and interesting messages, they can become annoying and leading to a negative perception of this feature.

#### H. Geolocation

The most significant element that should be added into a mobile app is geolocation, which allows users to identify their actual position by using a map that perfectly represents routes or other navigation data [27], [28]. Presently, GPS service for crises, or Location - based services, was among the most feature using in smartphone, making information sharing seamless and quickly [29]. Approximately 7 of the literatures analyzed [14], [16], [15], [17], [18], [12], [11] have built-in geolocation capabilities. Donor and centre locations are displayed on a visual map and optimised based on user real-time location data to identify the nearest one centre or donor. Since GPS services utilise users' personal information, harmful implications, as well as privacy and anonymity problems, might be crucial considerations that must be addressed to prevent data misuse.

RQ5. Who is the focused user of the system studies?

Based on 16 research articles reviewed, it was found that stakeholders who benefit from the use of charity applications are social organizations, donors, the needy, the community, authorized entities and patients. What can be seen from this study, researcher [8], [10], [11], [12], [26], [14], and [22] develop applications that focus on its main users involved in the system which are donors and charity centers and [15] develop similar systems but require an admin to manage donations from donors to social organizations before being sent to the receiver while [21], [23] involved with community. Researchers in [16] focus on the main users in the system which are donors and the community in blood finding. The author in [17] develops a blood finder between the donor and the authorizing party which is the hospital. As [16] in contrast to researchers [19], [18] [20] focused on connecting users and engagement between donors, admin, authorize the entity involved and the patient. Researchers in [9], [18] and [23] focused on the main users involved are donors and the needy who interact directly without an intermediary. As examined, the user is the component that determines how the user can communicate with the developed application. Therefore, indicating the system created is necessary to determine who will later be a user or users in the system in accordance with the system created.

#### IV. CONCLUSION AND FUTURE WORK

This literature review aims to systematically examine the extent to which charity application development and practice activities in the 2017-2021 range are based on the criteria exclusion and inclusion, 16 research articles were used for review. This literature review was carried out which is an approach to identify, collect and obtain information based on research question. Sixteen (16) articles were re-analyzed using five elements: the year, the article was published; the area of charity app develops that became the main focus of this research; the method used to conduct this study; the platform developed from the research; and the proposed feature and users who would use the system. In this finding, numerous academics have undertaken research to development of charity apps such as food donation, charity donations center, donation handle processing, crowdfunding and the search for blood donors. In literatures review resulted donation to charity center is the most discussing the topic. Furthermore, several methodologies are used in developing the apps, including structure method which consists of model waterfall, spiral, RAD, prototype, and object-oriented method. In this study, structured methods are the most commonly used and three platforms were discovered found in the literatures studies which mobile application development is dominance over web application and hybrid. From the aspects of functional, majority of apps installed include mobile features that could enhance the donation experience of the donors. Authentication, Notifications and Geolocation are the most interesting functionalities found in these studies. Research on system users, there are various users on each system developed, and it was discovered that donor users and charitable organizations are more in emphasis in the study. As a follow-up to this research, it is recommended in the form of suggestions, namely the need for further development and research on the application of charity in carrying out human activities and will be very useful during disasters or when times cannot be avoided.

#### ACKNOWLEDGMENT

This research was supported by Universiti Tun Hussein Onn Malaysia (UTHM) through Tier 1 Grant (Vot H937). The authors wish to thank Institute of Computer Science & Digital Innovation (ICSDI), CERVIE UCSI University which made this research endeavor possible.

#### REFERENCES

- [1] A. Razzaq, S. A. Asmai, M.S Talib, N. Ibrahim and Ali A. Mohammed. "Cloud ERP in Malaysia: Benefits, challenges, and opportunities." *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 9, no. 5, 2020.
- [2] Karyono, G., Ahmad, A., Asmai, S.A. "Survey on nudity detection: Opportunities and challenges based on 'awrah concept in islamic shari'a", *Journal of Theoretical and Applied Information Technology*, 95(15), pp. 3450–3460, 2017.
- [3] S. N. Wahab, and N. A. M. Radzi, "An Inquiry on Knowledge Management in Third-party Logistics Companies", *Int. J. of Business Innovation and Research*, 24(1), 124-146, 2021.
- [4] N. Bahar, S. N. Wahab, N. D. Ahmad, "Understanding challenges faced in online teaching and learning among Malaysian universities' instructors during COVID-19 pandemic", *Proceedings of the International Conference on e-Learning, ICEL, 154-157, 6th International Conference on e-Learning, ICEL 2020, 6-7 December 2020*, Sakheer, Bahrain. doi:10.1109/econf51404.2020.9385474. 2020.
- [5] D. Budgen and P. Brereton, "Performing systematic literature reviews in software engineering," *Proc. - Int. Conf. Softw. Eng.*, vol. 2006, no. December 2014, pp. 1051–1052, 2006, doi: 10.1145/1134285.1134500.
- [6] A. Jalil, M. Beer and P. Crowther, "Pedagogical Requirements for Mobile Learning: A Review on MOBIlearn Task Model", *Journal of Interactive Media in Education*, vol. 2015, no. 1, pp. 1-17, 2015. Available: 10.5334/jime.ap.
- [7] M. B. Aliyu, "Efficiency of Boolean Search strings for Information Retrieval," *Am. J. Eng. Res.*, vol. 6, no. November, pp. 216–222, 2017.
- [8] C. Dema, S. Zangmo, V. Mongar, and K. Dema, "Developing Charity Web Application to Eradicate Poverty in Bhutan," no. May, pp. 602–608, 2017.
- [9] A. Singh and D. S. Sharma, "Implement Android Application for Book Donation," *Proc. Int. Conf. Intell. Eng. Manag. ICIEEM 2020*, pp. 137–141, 2020, doi: 10.1109/ICIEEM48762.2020.9160283.
- [10] S. N. Wahab, Y. M. Loo, and C. S. Say, "Antecedents of Blockchain Technology Application among Malaysian Warehouse Industry", *Int. J. of Logistics Systems and Management*, 37(3), 427-444, 2020.
- [11] S. Belekar, R. Rajput, and K. Gharat, "Mobile Application for Donation of Items," vol. 1, no. 4, pp. 1–6, 2021.
- [12] R. I. A. Pribadi, A. Pambudi, and Ardiansyah, "EDonation Android Application for Used Goods Donation using Location-based Service," *J. Phys. Conf. Ser.*, vol. 1751, no. 1, 2021, doi: 10.1088/1742-6596/1751/1/012037.
- [13] S. A. B. Samud and S. Sulaiman, "Preloved Donation Mobile Application," pp. 1–5, 2017.
- [14] P. Lanerolle, S. Rathnayaka, H. Rupasinghe, and S. Madhushanka, "Donate . lk: A smart donation handling system," *2018 Natl. Inf. Technol. Conf.*, vol. National I, pp. 1–6, 2018.
- [15] S. R. Shelar, S. R. Salve, and A. S. Kedari, "A Smart Platform for Donation Handling," no. 5, pp. 42–45, 2020.
- [16] C. Olipas and Villanueva EM., "Dug-Uhay: A Blood Donor Finder Application," *Int. J. Trend Sci. Res. Dev.*, vol. 4, no. 1, pp. 757–762, 2019, [Online]. Available: www.ijtsrd.com.
- [17] G. Muhammad, H. Asif, F. Abbas, I. Memon, and H. Fazal, "An ERP Based Blood Donation Management System for Hospital and Donor," *Sukkur IBA J. Emerg. Technol.*, vol. 3, no. 1, pp. 44–54, 2020, doi: 10.30537/sjet.v3i1.542.
- [18] S. Ahdan and S. Setiawansyah, "Android-Based Geolocation Technology on a Blood Donation System (BDS) Using the Dijkstra Algorithm," *IJAIT (International J. Appl. Inf. Technol.)*, vol. 5, no. 01, p. 1, 2021, doi: 10.25124/ijait.v5i01.3317.
- [19] L. Sumaryanti, S. Suwarjono, and L. Lamalewa, "E-Blood Bank Application For Organizing and Ordering The Blood Donation," vol. 1, no. Icast, pp. 754–758, 2018, doi: 10.2991/icst-18.2018.153.
- [20] M. Nabil, R. Ihab, H. El Masry, S. Said, and S. Youssef, "A Web-based blood donation and Medical Monitoring System Integrating Cloud services and Mobile Application," *J. Phys. Conf. Ser.*, vol. 1447, no. 1, 2020, doi: 10.1088/1742-6596/1447/1/012001.
- [21] A. Asfarian, R. P. Putra, A. P. Panatagama, Y. Nurhadryani, and D. A. Ramadhan, "E-Initiative for Food Security: Design of Mobile Crowdfunding Platform to Reduce Food Insecurity in Indonesia," *2020 8th Int. Conf. Inf. Commun. Technol. ICoICT 2020*, 2020, doi: 10.1109/ICoICT49345.2020.9166180.
- [22] W. Lee, H. Yee, and N. Rahim, "Decentralized Application for Charity Organization Crowdfunding using Smart Contract and Blockchain," vol. 2, no. 2, pp. 236–248, 2021.
- [23] C. Varghese, D. Pathak, and A. S. Varde, "SeVa: A Food Donation App for Smart Living," *2021 IEEE 11th Annu. Comput. Commun. Work. Conf. CCWC 2021*, pp. 408–413, 2021, doi: 10.1109/CCWC51732.2021.9375945.
- [24] R. Kolandaisamy, S.L. Lie, I. Kolandaisamy, A.B. Jalil and G. Muthusamy, "The impact and effectiveness of e-wallet usage for Malaysian male and female", *Specialusis Ugdymas*, vol. 1, no. 43, pp. 4102-4109, 2022.
- [25] Asmai, Siti Azirah, Rosmiza Wahida Abdullah, Mohd Norhisham Che Soh, Abd Samad Hasan Basari, and Burairah Hussin. "Application of

- multi-step time series prediction for industrial equipment prognostic." *2011 IEEE Conference on Open Systems*, 273-277. 2011.
- [26] N. Bahar, S. N. Wahab and M. Rahman, "Impact of knowledge management capability on supply chain management and organizational practices in logistics industry", *VINE J. of Information and Knowledge Management Systems*, 51(5), 677-692, 2021.
- [27] Salleh, M. S., S. A. Asmai, H. Basiron, and S. Ahmad. "Named entity recognition using fuzzy c-means clustering method for Malay textual data analysis", *Journal of Telecommunication, Electronic and Computer Engineering (JTEC)*, vol. 10, no. 2-7, 121-126, 2018.
- [28] S. N. Wahab, R. Sham, A. A. A. Hussin, S. Ismail, and S. D. Rajendran, "Urban Transportation: A Case Study on Bike-Sharing Usage in Klang Valley", *Int. J. of Supply Chain Management*, 7(5), 470-476, 2018.
- [29] Asmai SA, Abidin ZZ, Basiron H, Ahmad S. An intelligent crisis-mapping framework for flood prediction. *Int. J. Recent Technol. Eng.* 8(2),1304-10. 2019.

# A Study on the Effect of Digital Fabrication in Social Studies Education

## Development of a Self-Learning Program for Creating 3D Educational Materials and Teaching Practice

Kazunari Hirakoso<sup>1</sup>

College of Engineering, Tamagawa University  
Tokyo, Japan

Hidetake Hamada<sup>2</sup>

College of Education, Tamagawa University  
Tokyo, Japan

**Abstract**—One of the learning methods that is increasingly being practiced in primary and secondary education is inquiry-based learning. This is not just a class to teach knowledge, but to practice activities to search for and discern the significance and essence of things. In social studies education, various trials and errors are being conducted, such as learning local history through fieldwork, and new approaches suitable for inquiry-based learning are being sought. In this study, as a new approach to social studies education, we developed a self-learning program that enables teachers to create original 3D educational materials using digital fabrication technology. We conducted an experiment in which students who wished to become social studies teachers participated in the program, created 3D educational materials, and taught a class using the materials. As a result, all the subjects who took the self-learning program could create 3D educational materials and give classes using them. The subjects' opinions suggested that practicing classes using 3D educational materials is effective for teacher education. This contributes to STEAM education, which has been spreading recently in the field of education, and this case study can be seen as a novel model.

**Keywords**—Digital fabrication; 3D educational materials; self-learning Program; social studies; STEAM education

### I. INTRODUCTION

In Japan, 3D printers are being promoted as school equipment in secondary education, making the use of 3D printers in school education more realistic. Usui et al. [1] report that the Ministry of Education, Culture, Sports, Science and Technology of Japan has added descriptions of 3D printers to its guidelines for developing educational materials for junior high schools and that 3D printer are now included in the curriculum guidelines for junior high-school technology and home economics courses and senior high school art, information, and industrial arts courses. There are already examples of 3D printers in science and technology education classes. Kurita et al. [2] use 3D topography created with a 3D printer as an aid in science classes to get an overview of the terrain, and they conduct classes that include on-site observation and fieldwork. Kadota et al. [3] incorporated the use of 3D CAD and 3D printers in a technology class to develop a radio-controlled car using a microcomputer, allowing students to learn multiple contents. Then, Muramatsu et al. [4] practiced and tested a learning curriculum for students

in a teacher training program for technology courses to build a model for introducing digital fabrication into school education. These precedents are probably due in part to the fact that digital fabrication is well matched to the content of these subjects.

However, the use of 3D printers in the classroom has been limited to a few subjects, and there are few practical examples, especially in social studies. This is probably because social studies teachers are far removed from technologies such as 3D printers and cannot imagine using them in social studies classes. Even if a 3D printer is close at hand, the opportunities that are not used well will limit the scope of inquiry-based learning.

In this study, we will develop a program that can practice digital fabrication, including 3D printers, in social studies education. It is a self-learning program that enables even beginners to create educational materials with a 3D printer and practice teaching with 3D educational materials. After class practice, we examine whether 3D educational materials are effective for social studies.

### II. PROGRAM DESCRIPTION

#### A. Previous Work

In developing the content to be used in this study, the teaching methods in e-learning in CAD were considered. Ahmed et al. [5] implemented a method for undergraduate students to demonstrate 3D modeling in an architectural 3D CAD class. As a result, students' motivation to learn and class understanding improved. Bodein et al. [6] indicated that in their survey of e-learning in CAD, they differentiated three types of instructional methods: awareness training (how to use the software), full training (how to operate the various functions), and performance support. In this study, we used teaching-by-demonstration and learning differentiated into steps.

#### B. Identification of the Teaching Content for beginner by Interview

The purpose of this program is to assist beginners in creating educational materials in history and geography through self-learning. We interviewed three teachers who teach digital fabrication to get an idea of the content needed for



beginners. All teachers have over eight years of experience teaching 3D CAD modeling and digital fabrication using 3D printers, CNC Milling Machine, etc. In the interview, we asked them to specify what beginners find difficult in the way of 3D model creation and digital fabrication. Furthermore, we found that beginners had difficulties when installing the software, so we made the program learnable from this point on.

C. Skills Required for Creating 3D Educational Materials and Self-Learning Program Contents

We identified the items required for the self-learning program. The content was designed so that beginners can learn step by step. Table I shows the skills needed to create 3D educational materials as a self-learning program. The process of creating 3D educational materials is divided into three steps, with each process consisting of detailed steps.

D. Selection of Contents of Social Studies Class

An example of a classroom practice studied in this program is the influence of topography on Japanese history. The 3D educational materials created by this program are modeled using a 3D printer with 3D topographic data provided by the Geospatial Information Authority of Japan (GSI) [7]. For more effective use of 3D educational materials, they can be used along with other educational materials.

E. Creation of Self-Learning Program

This self-learning program is built as a web system consisting of videos and an e-textbook. It consists of two parts. One is a method for creating 3D teaching materials composed mainly of videos, and the other is a classroom case study composed mainly of images and text. The 3D modeling software used was Fusion360 by Autodesk, Inc. [8].

Fig. 1 shows the screen where the 3D CAD operation method is explained. The instructor indicates the points to pay attention to in the voice while operating the software.

TABLE I. SELF-LEARNING PROGRAM CONTENTS FOR CREATING 3D EDUCATIONAL MATERIALS

Item No.	Step No.	Contents
1. What is 3D CAD and how to install it	1	What is 3D?
	2	What 3DCAD can do
	3	How to install 3DCAD and initial setup
2. 3D CAD Exercises	1	How to use key command operations
	2	2D sketching
	3	Geometric constraints
	4	3D modeling (Parametric modeling)
	5	Conversion to manufacturing data format
3. Features and Usage of Digital Machines	1	What is Digital Fab?
	2	Types of 3D printers and features of modeling
	3	Use 3D printer
	4	Features of CNC Milling Machines
	5	Use CNC milling machine
	6	Which machine to choose?

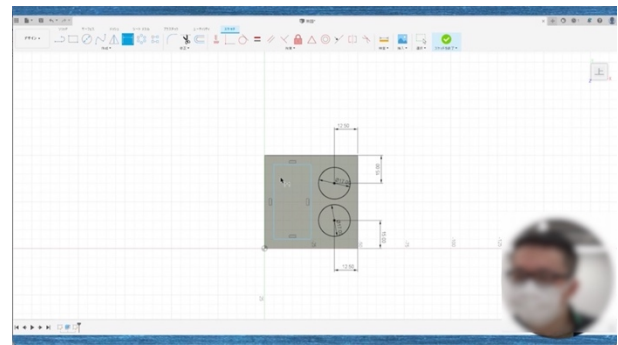


Fig. 1. An example of a self-learning program screen when explaining the operation of 3D CAD..

Fig. 2 shows an example of a class using 3D educational materials. The professor of social studies education explains key issues regarding specific classroom practices using 3D educational materials. The 3D educational materials created for the class are described with images and a written description of how to put them into practice.



Fig. 2. An Example of an Explanation of Key Issues in a Class using 3D Educational Materials.

Fig. 3 shows a mock class using 3D educational materials. The content is structured so that it can be implemented in class time.



Fig. 3. An Example of a Class using 3D Educational Materials.

III. EXPERIMENT

The following experiment was conducted over a total of eight weeks.

A. Self-Learning Program for Creating 3D Educational Materials

1) Overview of experiment: In this experiment, we asked subjects to take all the contents of the self-learning program in the correct order, up to the point where they completed the 3D educational materials. We investigated whether the subjects

could produce 3D educational materials without any problems after attending this program. After the experiment, subjects were asked the questions shown in Table II.

TABLE II. QUESTIONNAIRE FOR SELF-LEARNING PROGRAM FOR CREATING 3D EDUCATIONAL MATERIALS

Question	Level (1/2/3/4/5)
Q1. Were the self-learning program easy to understand?	1 – not at all 5 – very much
Q2. Were the speed of the explanation of the 3D CAD operation by video adequate?	1 – not at all 5 – very much
Q3. Were the content of the self-learning program sufficient for you to operate the 3D CAD?	1 – not at all 5 – very much
Q4. Did you find it stressful to understand the content of the self-learning program?	1 – very much 5 – not at all
Q5. If you were doing new class content, would you use a self-learning program like this one?	1 – not at all 5 – very much

2) *Subjects*: The subjects of this program are 20-year-old students: three males and three females, who aspire to become social studies teachers. Their PC skills are limited to daily use of Office software (Word, Excel, PowerPoint), and this is their first experience using specialized software such as 3DCAD.

3) *Experimental conditions*: Subjects took a self-learning program using their PCs. From the software installation to 3D data creation, the subjects used their PCs, and 3D educational materials were created at the Makers Floor (a fab facility equipped with 3D printers, CNC milling machines, and other digital fabrication equipment) located at Tamagawa University.

Fig. 4 shows the Form 2 [9] (Formlabs Inc.) 3D printer used in this study, which uses SLA printing: stereolithography, in which resin is cured by laser exposure for modeling. The software PreForm[10] (Formlabs Inc.) was used to convert 3D data (STL) to modeling data. The material used was Grey Resin. Fig. 5 shows an example of 3D educational materials. Fig. 6 shows the MDX-40A[11] (Roland DG Corporation) CNC-milling machine used in this study. The software SRP Player [12] (Roland DG Corporation) was used to convert 3D data (STL) to modeling data. The material used was chemical wood (SANMODUR MS-E). Fig. 7 shows an example of 3D educational material produced by a CNC milling machine.



Fig. 4. 3D Printer “Form 2 (Formlabs Inc.)”.



Fig. 5. An Example of 3D Educational Materials Modeled by 3D Printer.



Fig. 6. CNC Milling Machine “MDX – 40A (Roland DG Corporation).”



Fig. 7. An Example of 3D Educational Materials Modeled by CNC Milling Machine.

### B. Self-Learning Program for Teaching Practice in Social Studies Classes

1) *Overview of experiment*: Subjects first took the entire “Teaching Practice in Social Studies Classes” part of the self-learning program using their PCs. Then, each subject set their theme and reviewed the 3D educational materials they wanted to use in that class. Since the self-study program was also a collection of examples of classes using 3D educational materials, the subjects referred to it as needed when considering the content of their classes. After the experiment, subjects were asked the questions shown in Table III.

TABLE III. QUESTIONNAIRE FOR SELF-LEARNING PROGRAM FOR TEACHING PRACTICE IN SOCIAL STUDIES CLASSES

Question	Level (1/2/3/4/5)
Q1. Were the self-learning program easy to understand?	1 – not at all 5 – very much
Q2. Were the speed of the explanation of the 3D CAD operation by video adequate?	1 – not at all 5 – very much
Q3. Were the content sufficient for you to understand the significance of using 3D educational materials in your classroom?	1 – not at all 5 – very much
Q4. Have you been fully briefed on how to plan for implementing 3D educational materials in your classroom?	1 – not at all 5 – very much
Q5. If you were doing new class content, would you use a self-learning program like this one?	1 – not at all 5 – very much

2) *Subjects:* The subjects of this program are 20-year-old students: three males and three females, who aspire to become social studies teachers. These are the same subjects who attended the self-learning program for creating 3D educational materials.

3) *Experimental conditions:* After attending a self-study program on how to create 3D educational materials, the participants created and re-created 3D educational materials and discussed class content and class development over six weeks. After that, a mock class using 3D educational materials was conducted.

#### IV. RESULTS

All the subjects (n = 6) who took the self-learning program could create 3D educational materials and could practice teaching a mock social studies class for junior high school students using the 3D educational materials. Each experiment was conducted using questionnaires and interviews. The averages for each question are shown in Fig. 8 and Fig. 9. The results indicated high average values for both programs.

##### A. Self-Learning Program for Creating 3D Educational Materials

The results show that most subjects learned without stress and felt a slight deficiency in the amount of instructional content. Additionally, the following statements were made in the free response section of the questionnaire.

- After the experience, I felt a sense of joy and accomplishment when I created it came to me.
- I was relieved to find that with the self-learning program, creating 3D educational materials was easier than I had expected, even for someone like me who is not excellent with ICT.
- I could create the product by myself because I was carefully taught how to install the software and how to create the product using 3D CAD.
- I used to think that making things was something you did with your hands, but after the course, I learned that there is more than one way to make things.

- Using 3DCAD, I could create teaching materials like a game like Minecraft and produce them right in front of my eyes. This was especially interesting and fun for me.
- Although this program is for teachers, I thought it would be more accessible if there was a format that would allow children to experience it as well.

##### B. Self-Learning Program for Teaching Practice in Social Studies Classes

The results show that all items have high values, with easy-to-understand explanations rated especially high. Additionally, the following statements were made in the free response section of the questionnaire.

- I felt the significance of the advancement of technology and its use.
- I thought it was the knowledge I needed to know as a teacher.
- I created a 3D topography. I felt that the advantage of 3D topography is that it provides a bird's-eye view of the area through the senses of sight and touch.
- I thought the creator of 3D educational materials needs to understand the characteristics of what he or she wants to create and output it.
- I thought that a great deal of knowledge about 3D educational materials is needed both in terms of hardware at schools and among teachers themselves so that they can be used as educational materials in many schools in the future.

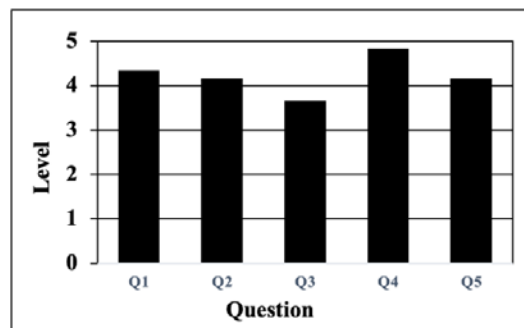


Fig. 8. Average Levels of Questionnaire Results of Self-Learning Program for Creating 3D Educational Materials.

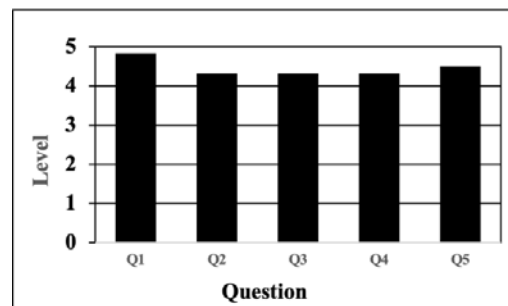


Fig. 9. Average Levels of Questionnaire Results of Self-Learning Program for Teaching Practice in Social Studies Classes.



## V. DISCUSSION

### A. Potential of Digital Fabrication in Social Studies Education

As a result of the self-learning program, first, the purpose of creating 3D educational materials was generally achieved; all the subjects completed the 3D educational materials, although there were a few who experienced difficulties in operating the 3D CAD system. This indicates that even beginners can create 3D educational materials if they follow the appropriate instructions since the program deals with technical content.

However, there are issues regarding the practice of teaching with 3D educational materials. Excerpts from the actual evaluation comments are provided below.

- The content of what to create as 3D educational materials could not be decided.
- The lack of practical examples makes it difficult to realize in class and limits the scope of the concept.
- Without the support of social studies discipline-based epistemological approach, it is difficult to know what to express using 3D educational materials.

Many comments on classroom practice were about the implications of the class, such as what kind of 3D educational materials should be created and how to develop a class based on them.

However, many commented that they were deeply interested in the practice of teaching with 3D educational materials and wanted to practice it more. Participants who completed lessons using 3D topography using data provided by the Geospatial Information Authority of Japan showed great interest in the possibilities offered by these 3D educational materials.

These results indicate that the self-learning program developed in this study is effective in providing technological support, while improvements are needed in areas related to the "teacher's discipline-based epistemological approach" in classroom practice using 3D educational materials.

### B. Effectiveness of using 3D Educational Materials in Social Studies Education as Teacher Education

Consideration of the opinions of the subjects of the self-learning program suggests that to be able to create 3D educational materials optimized for inquiry-based learning and implement them in their classes, teachers themselves must be able to articulate "what they want to communicate to their students." For example, if a teacher wants to teach students about the geography of an area, "What kind of terrain can be created to highlight its features?" or "Will the students be able to notice the natural principles themselves?" ... etc., the teacher will have to deeply consider such questions, determine their perspective, set the task, and question their background knowledge. Additionally, empathy will be needed to imagine from the student's perspective.

Thus, the creation of 3D educational materials optimized for inquiry-based learning is beneficial not only for the

teacher's reflection but also for the acquisition of one's perspective and the formation of a new view of teaching that emphasizes a discipline-based epistemological approach. The self-learning program developed in this study will also enable teachers to create their tailor-made teaching materials, thereby realizing a learning environment that is optimized for everyone.

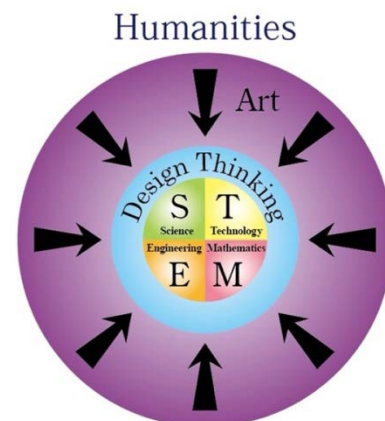
### C. Toward a Novel STEAM Education

After considering the results of the self-learning program and the process of implementation, we considered that the process for teachers to create effective 3D educational materials can be explained by the following new model, which includes design thinking in STEAM education. Fig. 10 shows the novel model. This is tentatively called the Hamada - Hirakoso model.

First, let us explain from outside the model illustration: we believe that the seeds of ideas for creating 3D educational materials are many across broad humanities. This means that the range of ideas itself would be limited if there is no intellectual activity such as consciously ingesting knowledge regularly.

However, knowledge is not enough to generate ideas. It is necessary to have one's perspective. This perspective can be viewed as an art of teaching. In this model, "one's perspective" and "sense of values" is "Art". It is only when one's perspective is established that one can decide what kind of educational material should be created, and having multiple perspectives allows one to consider various possibilities at the same time.

And the ability to access STEM expertise must create educational materials. This can be paraphrased as "the ability to implement. This ability of implementation is "Design thinking". The five steps of the process described in Design Thinking, "Empathy", "Define", "Ideate", "Prototype", and "Test", are used to link the seed of an idea to a solid basis. Additionally, through a process of trial and error, we searched for better educational materials. Without this process, good educational materials will never be completed. Design thinking will be used to access STEM, create educational materials, and implement classroom practices. We believe that teachers can use this model as a map for their daily practice.



© HAMADA H, HIRAKOSO K

Fig. 10. A Novel Model of STEAM Education.

## VI. CONCLUSION

In this study, we developed a self-learning program that enables students to practice digital fabrication, including modeling with a 3D printer, in social studies education.

In the experiment, subjects who were beginners in digital fabrication could create their own original 3D educational materials after taking this self-learning program. After considering the opinions of the participants in the self-learning program, creating 3D educational materials and teaching with 3D educational materials is a test of a teacher's ability. In other words, the results suggest that creating original 3D educational materials is effective for teacher education. This could be considered a novel model for STEAM education.

In the future, we intend to brush up on the program and expand the number of users by creating many examples of this program in practice. We would also like to verify whether the novel model we discovered was versatile enough.

## ACKNOWLEDGMENT

We would like to express our gratitude to the Makers Floor at Tamagawa University for their cooperation in the development of this program.

## REFERENCES

- [1] S. Usui, and Y Noborimoto, "An Examination of the Content Regarding 3D Printers in the Japanese Commentary to the Curriculum Guidelines", RESEARCH REPORT OF JSET CONFERENCES, Volume 2022, Issue 1, pp. 143-146, 2022.
- [2] K. Kurita, M. Morito, T Genda, Y Komatsu, M. Shibata, and H Shigematsu, "The Research and Development II for Teaching Material in Elementary and Junior high school Science Lessons: the lessons of geomorphology with solid model made by 3D printer", Bulletin of the Integrated Center for Educational Research and Training, Vol 50, pp 75-86, 2020.
- [3] K. Kadota, A. Inomata and H Nagashima, "Development of a Radio-controlled Car Using a Microcomputer Board in Junior High School Technology Education", Journal of the Japan Society of Technology Education, Volume 61, Issue 4, pp 297-304, 2019.
- [4] H. Muramatsu, K Kadota, H. Kawakubo, and D Doyo, "Proposal of digital craft introduction model at Faculty of Teacher Training", Proceedings of TENZ. ICTE Conference Technology: An holistic approach to education, pp. 223-231, October 2017.
- [5] V. Ahmed, L. Mahdjoubi, X. Feng, and M. Leach, "The learning of CAD for construction: technical abilities or visual?", INTERNATIONAL JOURNAL OF IT IN ARCHITECTURE ENGINEERING AND CONSTRUCTION, 2, 7-18, 2004.
- [6] Y. Bodein, R Bertrand and C. Emmanuel, "CAD Teams Performance Empowerment and Evaluation by Using E-Learning Tools." DS 58-10: Proceedings of ICED 09, the 17th International Conference on Engineering Design, Vol. 10, Design Education and Lifelong Learning, Palo Alto, CA, USA, 24.-27.08, 2009.
- [7] Geospatial Information Authority of Japan, the Japanese GSI: <http://www.gsi.go.jp/>.
- [8] Autodesk Inc., Fusion 360: <https://www.autodesk.com/products/fusion-360/overview>.
- [9] Formlabs Inc., Form 2: <https://formlabs.com/3d-printers/form-2/>.
- [10] Formlabs Inc. , PreForm : <https://formlabs.com/software/#preform>.
- [11] Roland DG Corporation, MODELA MDX-40A: <https://www.rolanddg.com/ja/news/2004/041125-modela-mdx-40>.
- [12] Roland DG Corporation, SRP Player : <https://www.rolanddga.com/support/products/software/srp-player>.

# A Blockchain-based Model for Securing IoT Transactions in a Healthcare Environment

Mohamed Abdel Kader Mohamed Elgendy<sup>1</sup>, Mohamed Aborizka<sup>2</sup>, Ali Mohamed Nabil Allam<sup>3</sup>

Computer Science Department<sup>1,2</sup>

Information Systems Department<sup>3</sup>

Arab Academy for Science, Technology and Maritime Transport (AASTMT)

Cairo, Egypt

**Abstract**—A blockchain is a data structure that is implemented as a distrusted database or digital ledger. The transactions are saved to a block of transactions that is attached in turn to the blockchain after the verification process, in which each block in the chain contains a hash signature of the previous block in addition to the hash signature of the block itself. The blocks on the blockchain are chained as an immutable list using the proof-of-work procedure, where there is no way to alter or delete an attached block due to the strict security policy used for structuring the chain of blocks. Each node holds a copy of the blockchain in which the miners take the responsibility of verifying and attaching blocks to the blockchain. The Ethereum blockchain introduced the smart contract which holds logic to be processed once the contract is established. These smart contracts are developed via the Solidity programming language. This proposed paper exploits the Ethereum blockchain along with smart contracts as the base technology for implementing the proposed blockchain-based model. The paper aims to develop a multilayered blockchain-based model, in which the blockchain model is set up on a private blockchain Ethereum network where the nodes share the electronic medical records (EMR) among the P2P (peer-to-peer) network that will be used to secure the IoT medical transactions. Solidity smart contract, introduced by Ethereum, is deployed to handle the EMR “open-query-transfer” operations on the private network, whereas the miners are responsible to validate the transactions. Finally, the research conducts a performance analysis of the Ethereum network using the Ethereum Caliper, considering several performance factors, which are: Maximum Latency, Minimum Latency, Average Latency, and Throughput.

**Keywords**—Blockchain; ethereum; electronic medical records (EMR); iot secure transactions; smart contracts; proof-of-work

## I. INTRODUCTION

The blockchain technology has significantly contributed to putting an end to the interoperability challenges found in current and legacy healthcare IT systems, enabling individuals, healthcare service providers, healthcare entities, and medical institutions to securely share electronic healthcare sensitive data. Blockchains can enhance communication efficiency and increase security over the network, as the potential for using blockchain in healthcare is to overcome the challenges related to data security, privacy, sharing and storage [6], as well. It can also be applied in many software domains including financial and banking sectors, healthcare systems, and public services.

Although the blockchain-based models are increasingly used in modern software solutions, such models raised a significant number of challenges and objectives such as scalability, performance, processing speed over the network, data management on distributed nodes, and security breaches and attacks. Moreover, IoT network devices have been growing rapidly, as the number of installed IoT devices in the year 2022 is estimated to be 31 million devices. However, over the past few years, the blockchain model has become more stable and most of such issues and concerns have been resolved at most of the blockchain well-known platforms.

Thereby, as blockchains have become an excellent candidate to replace traditional transaction database systems, strict standards including acceptable behavioral guidelines must be laid out. These guidelines will facilitate the process of integrating the blockchain technology onto the healthcare domain systems, within the two blockchain main types: private and public blockchains. Thus, a blockchain network is basically either public or private, where a private blockchain is constructed for usage on a private network mainly used within a single entity such as a financial institution, for example. Generally speaking, blockchains are immutable, and thereby miners hold the responsibility of verifying the attached blocks to a blockchain on both private and public blockchains. But in the case of private networks, miners validate the blocks with a much stricter secured policy.

Miners create blocks of transactions, verify the blocks, and then attach every verified block to the blockchain. In blockchains, proof of stake or proof of work is being used to control the difficulty of the mining process. Thus, raising the complexity of the hash computations within the proof-of-work algorithm would increase the verification time of the newly attached block to the blockchain.

A block consists of five basic components: previous hash signature, nonce, transaction, timestamp, and the hash signature generated using proof of work or proof of stake. Data within the block attached to the blockchain is immutable; it is extremely difficult to be altered due to distributed nature of the blockchain structure. Furthermore, each block on the blockchain has a reference to the previous block hash signature, any change on the signature cannot be accomplished as in this case the hash has to be recalculated, and as a result, the blockchain will detect that change through the data verification process of the block.

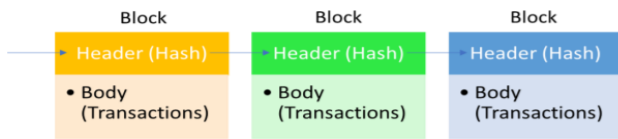


Fig. 1. Blockchain Architecture [16].

A block, which is chained onto a blockchain, holds transactions. As shown in Fig. 1, a block data structure consists of a block header and a block body. The block header is formed using multiple parameters which may vary based on the blockchain network provider. For the sake of integrity and verification, the header contains the parent hash signature which points to the previous block, in addition to the hash signature of the current block. The block body holds the block transactions. The block size is defined by the blockchain network provider, as this may vary based on the blockchain network provider.

The natural properties of the blockchain technology can be used to face the challenges mentioned previously, (1) permission-based blockchain networks to enable granular access control for medical records can be achieved by supporting granular-level access mechanisms; (2) blockchain-supported smart contracts enable patient-centric and transparent data sharing and control; (3) the blockchain distributed consensus mechanism overcomes the limitation of centralization; and (4) the immutable block preserves the integrity of data, which enables a blockchain to be verifiable and provable [1].

Accordingly, this paper presents a multilayered blockchain-based model for securing IoT transactions in healthcare traditional transaction systems to overcome the aforementioned issues concerning the dispersed and unified patients' medical records. Thus, the contribution of the proposed research is twofold:

- 1) Taking advantage of the blockchain technology as an immutable database for operating multilayered architecture flexible pattern, designed using clusters with embedded nodes to enforce flexible security level and approval permissions for medical records transactions on the whole blockchain.
- 2) Exploit the power of the Solidity programming language for developing highly secured smart contracts to handle the operation messages between the nodes from one side and the IoT devices and regular PC devices on the other side.

Therefore, the proposed multilayered blockchain model develops an approach for maintaining and managing patient medical records assembled in clusters with the usage of blockchain-based systems. The presented model uses aggregation (i.e., clusters) on the level of networks via (1) the Network ID, as each network represents a mining facility with its separate miners; (2) private blockchains on the cloud services provider "AWS" with the aim of creating the blockchains, the nodes, in addition to the miners; (3) Ethereum as an open-source blockchain-based distributed computing application platform. Additionally, the efficiency and performance of smart contracts are measured with the

sophisticated certified tool "Caliper" which is used for measuring the core functions of the smart contract: Open, Query, and Transfer.

Experiments that are conducted to test the blockchain difficulty configuration, in addition to the number of total miners on the blockchain, shall prove the capability of the proposed model to work with high-security complexity, and with medium or low-security complexity, as well.

This paper is organized as follows: Section II will review the blockchain-based smart home architecture. Besides, it will compare the currently used dispersed healthcare models based on both, relational and NoSQL database architectures. Consequently, Section III will present two main contributions, first is designing the blockchain-based model used for securing and managing data immutable storage for healthcare multi-layered medical systems. Secondly, exploiting the smart contracts with solidity for managing the health organization transactions along with the verification of these transactions. Section IV deals with Implementation of model while Section V presents the experiments that demonstrate the effectiveness of the proposed architecture. At the end of the paper, Section VI will provide a discussion of open problems and will lay out a direction for the future work that could be done.

## II. RELATED WORK

Dorri, et al. [2] presented a blockchain-based smart home architecture, shown in Figure 2, which has been used as a core reference in the process of designing our proposed model. This BC-based smart home architecture consists of three tiers, the smart home, the overlay, and the cloud storage, in which communication within these tiers is carried out using block transactions. The smart home consists of IoT devices, local IL, and local storage as demonstrated in Figure 2; overlay is a P2P network with distributed capabilities in addition to cloud storage groups based on identical unique block numbers, where SHM has been used for authentication.

### A. Blockchain Systems Versus Traditional Systems

Table I draws a comparison between the features of the blockchain-based systems and the remote patient monitoring system which depends on traditional communication and data storage methods, such as relational databases and cloud computing [3].

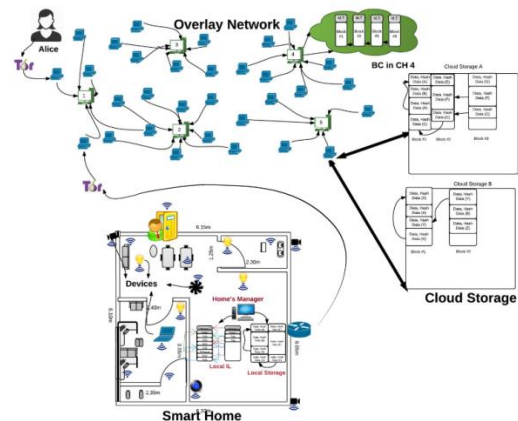


Fig. 2. BC-Based Smart Home.



TABLE I. COMPARISON BETWEEN TRADITIONAL AND BLOCKCHAIN-BASED SYSTEMS

Factor	Traditional Systems	Blockchain-based Systems
Confidentiality	Security level based on the configuration which may vary	High level of security
Availability	Must be manually configured	High service availability
Immutability	Data is exposed for manipulation	Immutable; the attached block cannot be altered or deleted
Traceability	Manually configured with a complex configuration	Traceable
Speed	Depends on network speed and hardware configuration along with the data source provider engine	May vary based on the blockchain verification process

A blockchain-based system runs on a P2P network of computers where each node on the network has an identical copy of the blockchain. Blockchains types can be classified as public, private, or hybrid blockchains.

1) Public blockchain: it was first implemented by Bitcoin and other cryptocurrencies, and it has contributed significantly to the distributed ledger technology (DLT) structure. Issues due to centralization are handled with DLT as it distributes data throughout a P2P network rather than storing it in a single location. Because of its decentralized nature, it forces methods of authentication.

2) Private blockchain: it is set up on a closed private network or controlled by a single entity. Functionality goes on the same basis regarding connectivity and decentralization; however, is substantially smaller.

3) Hybrid blockchain: it includes private and public blockchain characteristics. It allows the creation of a private permission-based system along with a public permissionless system, in addition to regulations for access to specific data on the blockchain [4].

### B. Smart Contracts using Solidity

A smart contract can be defined as a piece of code that lives on a blockchain and is then executed automatically when one or more conditions are met. In the case of the Ethereum blockchain, smart contracts are implemented via the “Solidity” object-oriented language, in which users can execute the smart contracts through an application binary interface [6]. This property enables entities to perform their job functionalities such as access management, request handling, and data transmission. Ethereum enhanced the communication between a patient and a physician, as sharing medical prescriptions with the patients became much faster and easier. Accordingly, patients share their historical treatment data with doctors in a fast and accurate manner [5].

### C. Comparison of the Data Management Mechanisms

Traditional legacy medical records are paper-based medical records (PMR), making it very difficult to keep track of a patient’s health history. Thus, saving historical data in such a way will cause data loss in addition to increasing the potential of inaccurate historical data, which potentially may lead to maltreatment. This serious issue has been faced by utilizing electronic medical records (EMR), the digital transformation of paper-based medical records. Electronic access to historical health records significantly improved the quality of treatment in addition to better disease diagnosis and preventive care [6]. Thereby, blockchain-based systems played an important role in modern healthcare solutions. Table II reviews and compares these main blockchain-based research exploited in the healthcare sector.

TABLE II. REVIEW OF THE BLOCKCHAIN-BASED RESEARCH IN THE HEALTHCARE SECTOR

Research	Blockchain Characteristics	Type(s) of Data	Merits
Castaldo & Cinque [7]	A private blockchain that does not rely on proof-of-work	EMR	Sharing E-health data across the EU via audit logging
Yue, et al. [8]	Private blockchain	EMR & PMR	Smart App to manage and share healthcare data
Patel [9]	A private blockchain that guarantees proof-of-stake	Medical Image Records	Securely sharing medical images
Fan, et al. [10]	Hybrid consensus mechanism based on practical byzantine fault tolerance (PBFT)	EMR	Secure sharing of healthcare data
Ji, et al. [11]	Proof-of-work	Patients’ Locations	Multilayer location sharing schema
Azaria, et al. [12]	Ethereum blockchain with proof-of-work	EMR	EMR management and sharing of healthcare data
Zhu, et al. [13]	Ethereum platform	EMR	Data management in the cloud environment
Genestier, et al. [14]	Hyperledger platform	Medical Records	Managing personal data in the e-health environment
Wang & Song [15]	Consortium blockchain	Medical Records	Coupling encryption and signature for robust security

### III. RESEARCH METHODOLOGY

In this section, we propose and develop an approach for maintaining and managing patient medical records assembled in clusters with the usage of blockchain-based systems. Basically, the proposed model uses clusters on the level of networks using the network ID, as each network represents a mining facility with its separate miners. The proposed approach exploits the cloud services provider “AWS” to create the blockchains, the nodes, and the miners. Also, the proposed model uses “Ethereum” as an underlying base technology for managing the blockchain operations; the Ethereum platform is widely used as an open-source blockchain-based distributed computing application utility.

Moreover, the proposed model uses smart contracts developed with the “Solidity” programming language, which will be first deployed with the address to the blockchain, and then executed using its current hexadecimal address (Open-Query-Transfer).

It is worth mentioning that “Solidity” exploits the hashing algorithm KECCAK-256, as an alternative to the NIST standardized SHA-3 hash function, to verify the chained blocks (i.e., proof-of-work). The algorithm is defined as:  $(m,n) = \text{POW}(H_n, H_n, d)$  where  $m$  is the mixHash,  $n$  is the nonce,  $H_n$  is the new block’s header,  $H_n$  is the nonce of the block header, and  $d$  is the DAG (is a large dataset). The mixHash is a hash that, when combined with the nonce, proves that this block has carried out enough computation.

Thus, the proof of work (POW) controls the level of difficulty of attaching a block to the blockchain, and as a result, increasing the difficulty level will make the process of formulating the hash which matches the target hash more complex and will eventually consume more time.

### A. Proposed Model

This research introduces a novel blockchain-based model for securing IoT transactions in the healthcare environment; the model that simulates the blockchain workflow on healthcare-based systems contains the following main components:

1) *Hospital / Clinic Miners*: miners are responsible for creating block transactions and attaching them to the blockchain.

2) *Transactions*: each transaction in the blockchain holds internal logic pertaining to the relevant smart contract. Once the transaction processing starts, the smart contract gets executed and the final output is conducted. There are three types of transactions within the proposed model: the “Open”, “Query”, and “Transfer” transactions, each of which is responsible for executing operations on the patient EMR file.

3) *Local Blockchain*: in each healthcare entity (hospital or clinic), a local private blockchain holds the blocks with its transactions, where each block has its own signature in addition to the previous block hash signature. The very first block in the blockchain is called the “genesis block”; it contains the setup configuration which controls the behavior of the blockchain in addition to controlling the security measures. The genesis block parameters will affect the process of attaching the new block of transactions, as the hashing algorithm will get harder based on the genesis block parameters, and thereby such configuration will directly affect the performance of the transactions’ processing.

4) *Global Blockchain*: as the healthcare sector usually consists of more than one medical entity, peering between these entities is established to sync the mining operations on different entities, in which such a peering process shall establish the global blockchain scope. The peering process is achieved via the “addpeer” request, and once the two blockchains are peered, the mining process is synced between the two blockchains, as demonstrated in Fig. 3. Upon successful peering request process, the mining operation will

be the responsibility of more than one miner, as each blockchain has one miner, each of which works at the same difficulty level value of the two blockchains.

5) *Overlay Network*: it enables a distributed functionality on the proposed model architecture. Clusters, which are a group of network nodes, are used to decrease network overhead.

### B. Proposed Model Design

Figure 4 illustrates the proposed model which clusters the network using the Network ID that represents a mining facility with its separate miners. The model also consists of private blockchains on the cloud services provider “Amazon Web Services (AWS)” for the sake of creating the blockchains and the nodes in addition to the miners, utilizing “Ethereum” as an open-source blockchain-based distributed computing application platform.

### C. Proposed Model Transaction Structure and Flow

On each entity, the private Ethereum blockchain is set up with the predefined setup configuration found on the Genesis block, where the mining process on the blockchain is controlled according to the setup configuration parameters on the genesis block. In each mining operation, a new block is created, and the transactions are included within the newly created block, and at the end, the block is attached and added to the chain.

```
> admin.addPeer("enode://61894226208310186f97d911b2206ae090fb9a59c21fc1f003d46c9
010e2a849ffcd1c2461adfd2acc5514f47a5a5a906f53e5fffb17231bcd6865abc0d2b906d054.188
.88.152:30303")
true
> DEBUG[05-08|21:56:37.396] Adding p2p peer                               name=Geth/v
1.8.27-stable-...
                                addr=54.188.88.152:30303 peers=1
DEBUG[05-08|21:56:37.396] Ethereum peer connected                       id=79935077a8
```

Fig. 3. Ethereum Blockchain Peering.

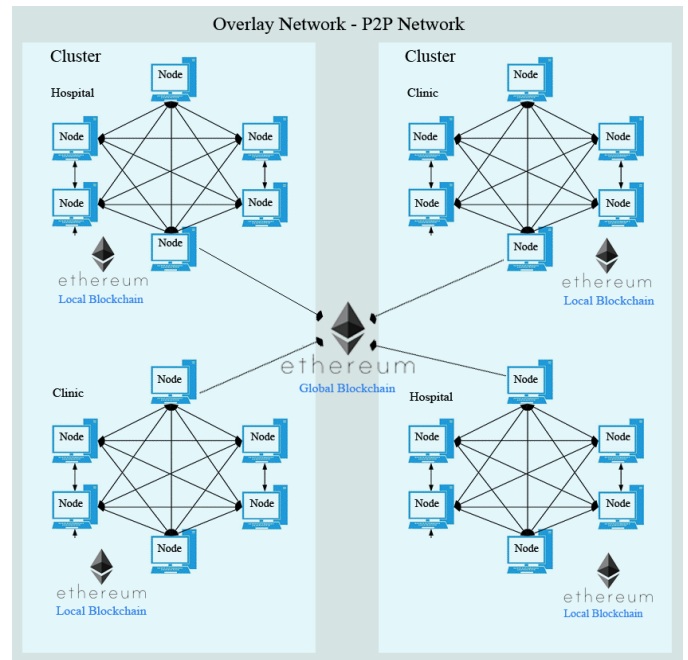


Fig. 4. Proposed Model Design.

Three main types of transactions are configured in the proposed model, namely, the “Open”, “Query”, and “Transfer” transactions, where each of them is executed via the smart contracts that contain the designated logic to be executed.

IV. IMPLEMENTATION

Fig. 5 is a flowchart that illustrates the steps of the addition and validation cycle in the proposed system.

A. Blockchain Initialization: The Genesis Block

The Genesis block contains the configuration of the private Ethereum network which will be used by all the miners, clusters, and nodes. The ChainId is used as the cluster identifier, where other configuration data inside the Genesis block will contain the following parameters/attributes:

- nonce: 64-bit string hash which, along with the mixHash, controls the amount of computation made for attaching the block to the blockchain.
- config: optional attribute which contains the ChainId unique identification of the private network; EIP150Block is used for fast sync, EIP155Block is used to reduce the probability of replay attacks, and EIP158Block controls how Ethereum clients handle empty accounts.
- timestamp: mainly used for verifying the order of the block within the blockchain.
- parentHash: a KECCAK 256-bit hash that points to the parent block.
- gasLimit: a scalar value that represents the limit of gas expenses of a single block in the blockchain.
- extraData: an optional parameter of 32 bytes at most, used for saving additional information if any.
- mixHash: a 256-bit hash which, together with the nonce, controls the level of computations used for verifying and attaching the block to the blockchain.
- coinbase: a 160-bit address, which is also called “etherbase”, holds all the successful mining operations amount.
- difficulty: a hexadecimal value that defines how hard it is to mine a block; the higher the value, the slower the mining process, since the mining operation will require more complex computations. Based on such difficulty value, hash computation is expected to run before obtaining a successful mining operation.
- alloc: optional parameters for predefining start balance on the mining account.

A typical example of the Genesis block structure:

```
{
  "timestamp": "0x5ca916c6",
  "nonce": "0x0000000000000042",
  "gasLimit": "0x2fefd8",
  "difficulty": "0x200",
```

```
"mixHash":
"0x0000000000000000000000000000000000000000000000000000000000000000",
  "coinbase":
"0xe1a7bcd0261e667651dc0e245d7b96e63c293d03",
  "number": "0x0",
  "config": {
    "chainId": 80,
    "eip150Block": 0,
    "eip155Block": 0,
    "eip158Block": 0 },
  "gasUsed": "0x0",
  "parentHash":
"0x0000000000000000000000000000000000000000000000000000000000000000",
  "alloc": {
    "0xe1a7bcd0261e667651dc0e245d7b96e63c293d03":
{
  "balance":
"0x2000000000000000000000000000000000000000000000000000000000000000"
} } }
```

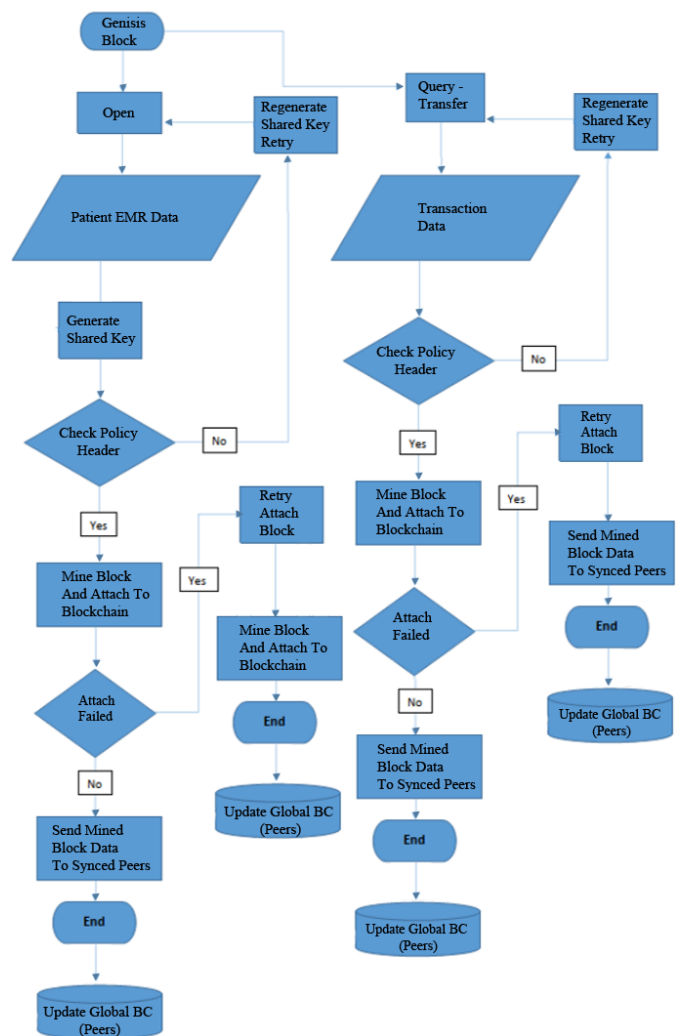


Fig. 5. Process Flow of the Proposed Model.







REFERENCES

- [1] M. K. Elghoul, S. F. Bahgat, A. S. Hussein and S. H. Hamad, "A Review of Leveraging Blockchain based Framework Landscape in Healthcare Systems," *International Journal of Intelligent Computing and Information Sciences*, vol. 21, no. 3, pp. 71-83, 2021.
- [2] A. Dorri, S. S. Kanhere, R. Jurdak and P. Gauravaram, "Blockchain for IoT Security and Privacy: The Case Study of a Smart Home," in *2017 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*, Kona, USA, 2017.
- [3] K. N. Griggs, O. Ossipova, C. P. Kohlios, A. N. Baccarini, E. A. Howson and T. Hayajneh, "Healthcare blockchain system using smart contracts for secure automated remote patient monitoring," *Journal of Medical Systems*, vol. 42, no. 7, pp. 1-7, 2018.
- [4] A. Haleem, M. Javaid, R. P. Singh, R. Suman and S. Rab, "Blockchain Technology Applications in Healthcare: An Overview.," *International Journal of Intelligent Networks*, vol. 2, pp. 130-139, 2021.
- [5] D. C. Nguyen, P. N. Pathirana, M. Ding and A. Seneviratne, "Blockchain for Secure EHRs Sharing of Mobile Cloud Based E-Health Systems," *IEEE Access*, vol. 7, pp. 66792-66806, 2019.
- [6] S. Khezr, M. Moniruzzaman, A. Yassine and R. Benlamri, "Blockchain Technology in Healthcare: A Comprehensive Review and Directions for Future Research," *Applied sciences*, vol. 9, no. 9, 2019.
- [7] L. Castaldo and V. Cinque, "Blockchain-based Logging for the Cross-border Exchange of eHealth Data in Europe," in *International ISICIS Security Workshop*, 2018.
- [8] X. Yue, H. Wang, D. Jin, M. Li and W. Jiang, "Healthcare Data Gateways: Found Healthcare Intelligence on Blockchain with Novel Privacy Risk Control," *Journal of Medical Systems*, vol. 40, no. 10, pp. 1-8, 2016.
- [9] V. Patel, "A Framework for Secure and Decentralized Sharing of Medical Imaging Data via Blockchain Consensus," *Health Informatics Journal*, vol. 25, no. 4, pp. 1398-1411, 2019.
- [10] K. Fan, S. Wang, Y. Ren, H. Li and Y. Yang, "Medblock: Efficient and Secure Medical Data Sharing via Blockchain," *Journal of Medical Systems*, vol. 42, no. 8, pp. 1-11, 2018.
- [11] Y. Ji, J. Zhang, J. Ma, C. Yang and X. Yao, "BMPLS: Blockchain-based Multi-level Privacy-preserving Location Sharing Scheme for Telecare Medical Information Systems," *Journal of Medical Systems*, vol. 42, no. 8, pp. 1-13, 2018.
- [12] A. Azaria, A. Ekblaw, T. Vieira and A. Lippman, "Medrec: Using Blockchain for Medical Data Access and Permission Management," in *2nd International Conference on Open and Big Data (OBD)*, Vienna, Austria, 2016.
- [13] L. Zhu, Y. Wu, K. Gai and K.-K. R. Choo, "Controllable and Trustworthy Blockchain-based Cloud Data Management," *Future Generation Computer Systems*, vol. 91, pp. 527-535, 2019.
- [14] P. Genestier, S. Zouarhi, P. Limeux, D. Excoffier, A. Prola, S. Sandon and J.-M. Temerson, "Blockchain for Consent Management in the ehealth Environment: A nugget for Privacy and Security Challenges," *Journal of the International Society for Telemedicine and eHealth*, vol. 5, pp. 1-4, 2017.
- [15] H. Wang and Y. Song, "Secure Cloud-based EHR System using Attribute-based Cryptosystem and Blockchain," *Journal of Medical Systems*, vol. 42, no. 8, pp. 1-9, 2018.
- [16] M. N. M. Bhutta, A. A. Khwaja, A. Nadeem, H. F. Ahmad, M. K. Khan, M. A. Hanif, H. Song, M. Alshamari and Y. Cao, "A Survey on Blockchain Technology: Evolution, Architecture and Security," *IEEE Access*, vol. 9, pp. 61048-61073, 2021.



# Study on Early Warning on the Financial Risk of Project Venture Capital through a Neural Network Model

Xianjuan Li

Hunan City University  
Yiyang, Hunan 413000  
China

**Abstract**—This paper aims to effectively reduce the financial loss of enterprises by accurately and reasonably making early warning of investment project risks. This paper briefly introduced the index system used for investment project risk early warning. It constructed a project investment risk early-warning model with a back-propagation neural network (BPNN) algorithm, and improved it with a genetic algorithm (GA) to solve the defect that the traditional BPNN is easy to fall into, over-fitting when reversely adjust parameters. An analysis was conducted on an electric power company in Hunan Province. Orthogonal experiments are performed to determine the population size and the number of hidden layers in the improved BPNN algorithm. The results showed that the improved BPNN algorithm had the best performance when the population size was set as 25 and the number of hidden layers was four; compared with support vector machine (SVM) and traditional BPNN algorithms, the GA-improved BPNN algorithm had better performance for early risk warning of investment projects. In conclusion, adjusting the parameters of a BPNN with a GA in the training stage can effectively avoid falling into over-fitting, thus improving the early warning performance of the algorithm; in addition, the improved BPNN has better early warning performance.

**Keywords**—Neural network; project investment; early risk warning; genetic algorithm

## I. INTRODUCTION

The rapid development of the economy has led to the emergence of various new enterprises, which further promotes the development of the economic market. In the process of enterprise development, the scale of an enterprise will be expanded, and during expansion, in addition to the existing profitable projects, the enterprise will also invest in other profitable projects [1] to further expand revenue and accelerate its development. However, there are few projects in the market that can make steady earnings, and more often than not, the projects available for investment carry different degrees of risk. Generally speaking, the higher the risk of an investment project is, the higher the ultimate return is, but the high risk of an investment project also means that the project is more likely to fail and lead to losses. The risk level of a project depends not only on the success probability of the project but also on the ability of enterprises to bear the losses after project failure [2]. When faced with the same project with a probability of failure, large enterprises that have more financial support than small

enterprises can still operate normally even if the project fails, while small enterprises may not be able to operate normally, i.e., small enterprises will take more risks when facing the project. Therefore, before investing in a risky project, an enterprise needs to make an early warning assessment of project risks in conjunction with its financial situation to minimize the loss of the venture investment. The early warning assessment of a venture usually requires the appropriate professional knowledge of the assessor, but the managers of enterprises generally do not have the relevant professional knowledge [3]. Relying excessively on expert experience and subjective judgment when making decisions on a venture will seriously affect early risk warning. Therefore, enterprises need a relatively perfect project venture capital model to objectively warn the risk of investment projects and guarantee the smooth operation of the projects. This paper briefly introduced the index system used for investment project risk early warning, constructed a project investment risk early-warning model with a back-propagation neural network (BPNN) algorithm, optimized it with a genetic algorithm (GA), and analyzed an electric power company in Hunan Province. The novelty of this paper lies in the use of the GA to adjust the parameters in the BPNN, avoiding falling into over-fitting when adjusting the parameters reversely. The organization of this paper is introduction, related works, the construction of the neural network-based early financial warning model, example analysis, discussion, and conclusion.

## II. RELATED WORKS

Zhu et al. [4] constructed a financial risk early-warning model based on the K-means clustering algorithm and found that the K-means clustering algorithm effectively avoided the negative subjective impact brought by artificially divided thresholds. Sun et al. [5] constructed a back-propagation neural network (BPNN)-based financial early-warning model, took mining listed companies as the research object, and found that the constructed early-warning model had a high prediction accuracy. Li et al. [6] introduced the L1 regularized support vector machine (L1-SVM) into the modeling of financial early warning systems as an effective feature selection technique and verified the feasibility of the technique in practical applications. Ouyang et al. [7] proposed a long short-term memory (LSTM) neural network under an attention mechanism for early warning of financial market risks. The final experimental

results showed that this neural network had good generalization ability and higher prediction accuracy compared with BPNN, support vector regression (SVR), and autoregressive integrated moving average (ARIMA) models. Qu et al. [8] put forward an improved kernel principle component analysis-based financial risk prewarning model for public hospitals, conducted experiments on the financial data of multiple public hospitals and listed companies, and verified the feasibility and effectiveness of the method. Ding [9] proposed to establish a fuzzy theory-based early risk warning management and intelligent real-time monitoring model system for financial enterprises and analyzed a listed company engaged in automobile sales. His study found that the use of fuzzy theory and modern network technology provided more accurate early warning and assessment of potential and apparent risks of financial enterprises. Feng et al. [10] constructed a BPNN-based enterprise financial risk prewarning model to predict financial crises and verified that the constructed prediction model could accurately predict the financial crises of the enterprises through predicting the finance of 200 manufacturing enterprises in 2018 and 2019. Qi et al. [11] proposed a variable precision rough set weighted k-nearest neighbor (KNN) network-based financial risk control algorithm and verified the algorithm's algorithm through experiments. Zhang [12] used fuzzy neural networks to warn the credit risk of financing platform loans, verified the effectiveness of the algorithm by example analysis, and gave relevant suggestions.

### III. NEURAL NETWORK-BASED EARLY FINANCIAL RISK WARNING

#### A. Constructing Early Risk Warning Indicators

Before warning financial risks of an investment project, it is necessary to build an indicator system that can determine the investment risks of the project, and these indicators will be used as input parameters of the project risk early-warning model [13]. Selecting early risk warning indicators generally follows the criteria of indicator criticality, data source accuracy, indicator relativity, indicator validity, and indicator simplicity. Indicator criticality means that the selected indicators are related to capital flow. Data source accuracy means that the selected indicators can be obtained from the data sources. Indicator relativity means that the selected indicators are relative indicators, ignoring the influence of company size as much as possible. Indicator validity means that the selected indicators need to be universal and valid. Indicator simplicity means that the selected indicators should be easy to calculate.

Different companies will invest in different types of projects due to their different positioning, and the risks possessed by different types of investment projects are also different. Therefore, only the general classification of indicators is given here, and the specific indicators will be given in the example analysis below. As shown in Fig. 1, the risk variables of project investment can be broadly classified into four types of risks: economic, technical, policy, and environmental risks. Project economic risk refers to the crisis faced by the investment project at the economic level. Project technical risk refers to the technical risk that arises during enterprise operation that can affect the project. Project policy

risk refers to the degree of influence that can be caused by laws and regulations in the project investment process. Project environmental risk refers to the degree of influence of the local environment during the operation of the investment project, and outdoor projects are more likely to be affected by the environment, which requires specific analysis in different cases [14].

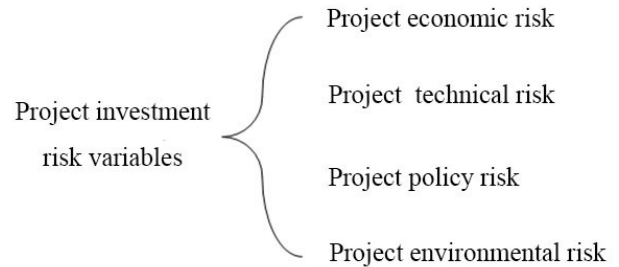


Fig. 1. General Classification of Early Risk Warning Indicators for Investment Projects.

#### B. Neural Network-Based Project Risk Early-Warning Model

The evaluation indicators of investment projects will be input into the input layer of the BPNN algorithm [15]; therefore, the number of nodes in the input layer depends on the number of indicators used to evaluate the risk of the investment project. The output layer outputs the evaluation results of the risk level of the investment project. The risk level is represented by 1, 2, 3, 4, and 5. "1" represents a low risk, and the larger the value is, the higher the risk is. The hidden layer is the core structure of the BPNN algorithm, and its number is decided according to the demand. Usually, the more the layers and nodes are, the deeper the law can be mined, and the more accurate the model prediction is, but it will increase the amount of computation [16]. The basic process of constructing an early warning model for investment project risk is as follows.

1) The indicator data of different investment projects are collected according to the constructed financial risk early warning indicator system, as well as the risk assessment results of corresponding projects.

2) The collected data are pre-processed to eliminate the abnormal data. The indicator data of investment projects are input into the input layer of the BPNN algorithm.

3) The indicator data in the input layer are calculated layer by layer in the hidden layer [17]:

$$a = f\left(\sum_{i=1}^n \omega x_i - \beta\right), \quad (1)$$

where  $a$  is the output of every layer,  $\beta$  is the adjustment term of every layer,  $f(\bullet)$  is the activation function [18], and  $\omega$  is the weight between layers.

4) The output result of the last hidden layer is passed to the output layer. The softmax function calculates in the output layer. The risk level of the investment project is output according to the calculation result.

5) The risk level of investment projects calculated by the BPNN algorithm is compared with the actual risk level obtained when collecting data, and the error between them is calculated. The cross-entropy [19] is used as the calculation error. The calculation formula of the error is:

$$E = -\sum_i y_i \cdot \ln p_i \quad (2)$$

where  $E$  is the error,  $i$  is the label serial number, which is the risk level,  $y_i$  is the judgment parameter [20], whose value is 1 when the actual risk level of the project is  $i$  and 0 otherwise, and  $p_i$  is the probability that the project risk level is  $i$  in the calculation result.

6) Whether the BPNN algorithm reaches the termination condition is determined. If it does, then the training ends, and the construction of the early risk warning model ends; if not, then the weight parameter in the hidden layer is reversely adjusted. The termination conditions for training the model are that the number of iterations reaches the preset maximum value or the calculation error converges to the preset threshold. The training is stopped when either of the above two termination conditions are met.

The adjustment of the weight parameter in the hidden layer is based on the calculation error and learning rate. The output result converges in the direction of minimum error through the calculation error and learning rate. In this reverse adjustment process, the learning rate is crucial as it controls the convergence speed of the algorithmic model, and it is usually a fixed value; however, in the practical application process, there are a large number of local minima in the nonlinear error surface, and once the model training falls into the local minima, it is difficult to get out, which will seriously slow down the convergence speed [21]. Therefore, the GA is introduced to adjust the weight parameter of the BPNN algorithm.

The process of training the improved BPNN algorithm using the GA [22] is as follows. Firstly, the chromosome population is generated for the GA. Every chromosome represents a parameter scheme of the BPNN algorithm, and every gene in the chromosome represents a parameter to be adjusted. Then, the parameter schemes represented by the chromosomes are substituted into the BPNN algorithm for forward computation according to steps (1)~(5) described previously to obtain the error. Then, whether the training should be terminated is determined. If not, the genetic operation is performed on the chromosome population, including crossover and mutation [23]. The crossover operation refers to exchanging the data on the same gene locus of two chromosomes according to the crossover probability, and the mutation operation refers to changing data at a single chromosome locus according to the mutation probability. The genetically manipulated population is substituted into the BPNN algorithm again to repeat steps ①~⑤ are repeated until the model reaches the termination condition.

## IV. EXAMPLE ANALYSIS

### A. Analysis Object

An electric power company in Hunan Province was taken as an example. The work that this power company can undertake includes power engineering survey, manufacturing, design, and sales. The company has a relatively good organizational structure. The shareholders' meeting is the highest authority of the company and appoints other departments. The board of directors is the representative department elected by the shareholders' meeting to manage the company's business operations, and the operating management layer established under the board of directors manages 11 departments.

The basic process for project investment is to bid and process the winning project. The company's management department does not have a set of strict evaluation procedures. The department manager expresses his investment intention first, and then the financial department decides whether the project can be invested in after a simple qualitative analysis. The whole process is highly subjective and nonstandard.

In addition to qualitative analysis that can determine the presence or absence of project risks, quantitative analysis is also needed to determine the level of project risks to help the management layer make more scientific and rigorous judgments [24].

### B. Project Risk Early Warning Indicator System

The project data required for the case study were collected from the information disclosed on the official website of the company. A field survey was conducted on the company to collect project analysis information such as project investment plans and cost reports that are publicly available.

These data were pre-processed before formal use, which was because the data volume that the company could provide was limited and some projects were suspected of fraud. Pre-processing supplemented some data and eliminated the part that could not be supplemented. After pre-processing, the total number of samples was 500, of which 300 were randomly selected as the training set and the remaining 200 as the test set.

Before establishing the improved BPNN-based early risk warning model, the corresponding early risk warning indicator system was established. This paper analyzed the early warning indicators of this electric power company by referring to the general classification of the indicators given in the previous section. There were thirty-four second-level indicators under the four first-level indicators of economy, technology, policy, and environment. Although more early warning indicators were good for prediction accuracy, the calculation volume was also larger. Some indicators had low relevance and would not affect the prediction even if ignored. Thus, the 34 second-level indicators were screened to eliminate those with low relevance to reduce the computational effort.

Table I shows the early warning indicators screened after the KMO test, Bartlett test [25], and regression analysis, the KMO test statistic was 0.736, which exceeded 0.7, and the Bartlett test statistic was 276.35. In addition, the p-values of all 13 indicators in Table I were less than 0.01.

TABLE I. THE SCREENED-PROJECT RISK WARNING INDICATOR SYSTEM

The first-level indicators	The second-level indicators	P-value	Kaiser-Meyer-Olkin test	Bartlett's test
Project economic risks	Project cost-income ratio	0.001	0.736	276.35
	Net income ratio	0.000		
	Asset turnover ratio	0.003		
	Turnover of account receivable	0.000		
	Asset-liability ratio	0.003		
	Liquidity ratio	0.002		
	Sales growth rate	0.000		
Preservation and appreciation ratio	0.000			
Project technical risks	Payback period	0.002		
	Construction technology	0.000		
	Construction management	0.000		
Project environmental risks	Project approval method	0.001		
	Social conduct	0.002		

C. Parameter Setting

A BPNN algorithm modified by the GA was used to construct a project investment risk early-warning model. The number of nodes in the input layer of the BPNN algorithm was set as 13 according to the feature number of indexes that need to be input, and the sigmoid function was used as the activation function in the hidden layer for fitting nonlinear laws. For the GA-improved BPNN algorithm, factors that affected its prediction performance also included the population size of the GA used for adjusting the weight parameter in addition to the number of hidden layers in its structure. Therefore, orthogonal experiments were used to test the performance of the improved BPNN algorithm with one, two, three, four, and five hidden layers under the genetic population size of 10, 15, 20, 25, and 30. The population size and the number of hidden layers with the best performance were selected and used for the subsequent comparison experiments.

To further test the early warning performance of the improved BPNN algorithm for project risks, it was compared with SVM and traditional BPNN algorithms. The parameters of the SVM algorithm are as follows. The sigmoid function was used as the kernel function for mapping features to a high-dimensional space to linearize nonlinear features as much as possible, and the penalty factor was set as one. The number of nodes of input and output layers and the number of hidden layers of the traditional BPNN algorithm were the same as those of the improved BPNN algorithm, the activation function was used as the activation function, and the learning step length was 0.02.

D. Experimental Results

Table II and Fig. 2 show the results of the orthogonal experiments for the population size and the number of hidden layers of the improved BPNN algorithm. First, it was noticed from Fig. 2 that the overall accuracy of the improved BPNN algorithm for investment project risk prediction increased as the population size and the number of hidden layers increased,

but the overall accuracy of the improved BPNN algorithm tended to be constant after the population size reached 25, and the overall accuracy of the improved BPNN algorithm also tended to be constant after the number of hidden layers reached four. However, comparing the average single-project time consumption under different population sizes and hidden layers in Table II, it was found that the average single-project time of the algorithm always increased as the population size and the number of hidden layers increased. In other words, increasing the population size and the number of hidden layers could increase the overall accuracy of the improved BPNN algorithm and also increase the prediction time, but after they increased to certain levels, the prediction time still increased, but the overall accuracy of the prediction tended to be constant. Therefore, the population size of the improved BPNN algorithm was 25, and the number of hidden layers was four.

Fig. 3 shows the test results of the early risk warning performance of SVM, traditional BPNN, and improved BPNN algorithms for investment projects. It was seen that the accuracy, recall rate, and F-value of the SVM algorithm for early risk warning of investment projects were 75.3%, 70.1%, and 72.6%, respectively; the accuracy, recall rate, and F-value of the traditional BPNN algorithm for early risk warning of investment projects was 91.2%, 82.1%, and 86.4%, respectively; the accuracy, recall rate, and F-value of the improved BPNN algorithm for risk warning of investment projects was 97.5%, 96.6%, and 97.0%, respectively. It was seen from Fig. 3 that the accuracy, recall rate, and F-value of the SVM-based early-warning model were the lowest, and those of the GA-based BPNN model were the highest.

TABLE II. AVERAGE TIME SPENT BY THE IMPROVED BPNN ALGORITHM WITH DIFFERENT POPULATION SIZES AND NUMBER OF HIDDEN LAYERS ON A SINGLE PROJECT

	One hidden layer	Two hidden layers	Three hidden layers	Four hidden layers	Five hidden layers
Population size 10	98 ms	110 ms	131 ms	163 ms	205 ms
Population size 15	121 ms	140 ms	162 ms	184 ms	213 ms
Population size 20	176 ms	195 ms	211 ms	234 ms	252 ms
Population size 25	223 ms	241 ms	264 ms	287 ms	303 ms
Population size 30	251 ms	272 ms	295 ms	316 ms	339 ms

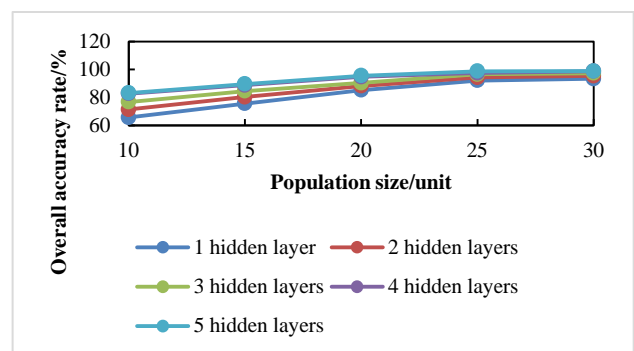


Fig. 2. Overall Accuracy of the Improved BPNN Algorithm under different Population Sizes and Number of Hidden Layers.

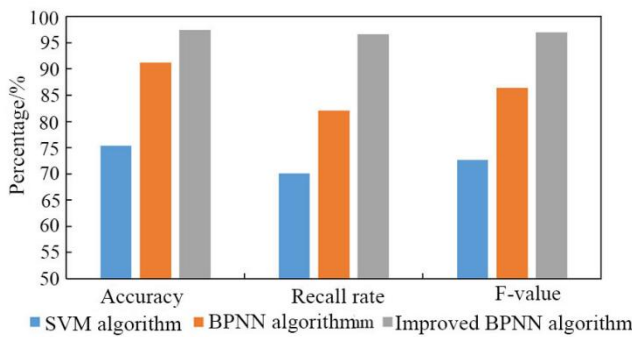


Fig. 3. Performance of Different Project Investment Risk Early-warning Models.

## V. DISCUSSION

Conventional enterprises will continue to expand their scales in the process of development, and they will make investments in different projects to obtain returns in the process of expansion. However, all project investments are risky. Usually, the greater the risk, the higher the return, but high risks also means that the probability of project failure. Once a project fails, the enterprise will suffer from huge losses, which is not conducive to its development. Thus, before investing in a new project, companies need to assess the risk to assist them in making decisions about the project investment. Traditional risk assessment is done manually, which is inefficient and subjective. In order to improve the efficiency of risk assessment and also to enhance the objectivity of the assessment results, intelligent algorithms are introduced into the early warning of project investment risks. This paper selects the BPNN algorithm, which can make a good fit to the nonlinear law, to warn the project investment risk and improved the traditional BPNN algorithm with the GA. Finally, an electric power company in Hunan province was analyzed, and the improved BPNN algorithm was compared with SVM and traditional BPNN algorithms. The experimental results have been shown in the previous section.

In the orthogonal experiments conducted by the improved BPNN algorithm on the genetic population size and the number of hidden layers in the BPNN, the early warning accuracy of the algorithm gradually increased but also stabilized as the population size and the number of hidden layers increased, and meanwhile the evaluation time also increased. The reason is as follows. The increase in the population scale made the parameters lead to more choices of parameters in the BPNN and increased the possibility of finding suitable parameters, and the increase in the number of hidden layers allowed the algorithm to fit the nonlinear law better, thus the early warning accuracy increased. However, both the increase in the population size and the increase in the number of hidden layers increased the computational effort of the algorithm, leading to an increase in computation time.

The comparison of the three early warning algorithms showed that the improved BPNN algorithm had the best performance, followed by the traditional BPNN algorithm, and the SVM algorithm had the poorest performance. The reason is as follows. Although the SVM algorithm could classify the risk level of investment projects relatively quickly and effectively,

the hyperplane it used was difficult to fit the nonlinear law; the traditional BPNN algorithm could fit the nonlinear law better, but the local minimum in the error surface in the training process could make the algorithm converge prematurely; after the improvement by the GA, the crossover and mutation operations adjusted the parameters to avoid falling into the local minimum, and the BPNN algorithm fully applied its nonlinear fitting to explore the laws, so it performed better in early warning.

## VI. CONCLUSION

This paper briefly introduced the indicator system used for investment project risk early-warning and used the BPNN algorithm to construct the project investment risk early-warning model. The BPNN algorithm was improved by the GA. An electric power company in Hunan Province was taken as a subject for analysis. The population size and the number of hidden layers in the improved BPNN algorithm were determined by orthogonal experiments. The results are as follows. (1) The increase in population size and the number of hidden layers in the improved BPNN algorithm improved the early warning accuracy; when the accuracy tended to be constant, the average time spent on early warning for a single project increased, so the final population size was set as 25, and the number of hidden layers was set as 4. (2) The accuracy, recall rate, and F-value of the SVM-based early-warning model were the lowest, those of the traditional BPNN algorithm-based model were higher, and those of the GA-improved BPNN algorithm-based model were the highest.

## REFERENCES

- [1] Q. Zhang, Q. Zhang, and D. Sornette, "Early Warning Signals of Financial Crises with Multi-Scale Quantile Regressions of Log-Periodic Power Law Singularities," *PLoS ONE*, vol. 11, pp. e0165819, 2016.
- [2] M. V. Klibanov, and A. V. Kuzhuget, "Profitable forecast of prices of stock options on real market data via the solution of an ill-posed problem for the Black-Scholes equation," *Papers*, vol. 32, 2015.
- [3] C. H. Huang, J. L. Yo, and A. P. Chen, "Using the Event Classifier System to Forecast the Stock Price: An Empirical Study of Taiwan Financial Market," *Appl. Mech. Mater.*, vol. 764-765, pp. 7, 2015.
- [4] Z. Zhu, and N. Liu, "Early Warning of Financial Risk Based on K-Means Clustering Algorithm," *Complexity*, vol. 2021, pp. 1-12, 2021.
- [5] X. Sun, and Y. Lei, "Research on financial early warning of mining listed companies based on BP neural network model," *Resour. Policy*, vol. 73, pp. 102223, 2021.
- [6] J. Li, Y. Qin, D. Yi, Y. Li, and Y. Shen, "Feature Selection for Support Vector Machine in the Study of Financial Early Warning System," *Qual. Reliab. Eng.*, vol. 30, pp. 867-877, 2015.
- [7] Z. S. Ouyang, X. T. Yang, and Y. Lai, "Systemic financial risk early warning of financial market in China using Attention-LSTM model," *North Am. J. Econ. Finance*, vol. 56, pp. 1-16, 2021.
- [8] M. Qu, and Y. Li, "Financial Risk Early-Warning Model Based on Kernel Principal Component Analysis in Public Hospitals," *Math. Probl. Eng.*, vol. 2021, pp. 1-7, 2021.
- [9] Q. Ding, "Risk early warning management and intelligent real-time system of financial enterprises based on fuzzy theory," *J. Intell. Fuzzy Syst.*, vol. 40, pp. 1-11, 2020.
- [10] Q. Feng, H. Chen, and R. Jiang, "Analysis of early warning of corporate financial risk via deep learning artificial neural network," *Microprocess. Microsy.*, vol. 87, pp. 1-8, 2021.
- [11] M. Qi, Y. Gu, and Q. Wang, "Internet financial risk management and control based on improved rough set algorithm," *J. Comput. Appl. Math.*, vol. 384, pp. 1-9, 2021.

- [12] J. Zhang, "Investment risk model based on intelligent fuzzy neural network and VaR," *J. Comput. Appl. Math.*, vol. 371, pp. 1-10, 2020.
- [13] W. X. Li, C. S. Chen, and J. J. French, "Toward an early warning system of financial crises: What can index futures and options tell us?," *Q. Rev. Econ. Financ.*, vol. 55, pp. 87-99, 2015.
- [14] S. P. Das, and S. Padhy, "A novel hybrid model using teaching-learning-based optimization and a support vector machine for commodity futures index forecasting," *Int. J. Mach. Learn. Cyb.*, vol. 9, pp. 97-111, 2018.
- [15] X. D. Zhang, A. Li, and P. Ran, "Stock trend prediction based on a new status box method and AdaBoost probabilistic support vector machine," *Appl. Soft Comput.*, vol. 49, pp. 385-398, 2016.
- [16] J. Patel, S. Shah, P. Thakkar, and K. Kotecha, "Predicting stock and stock price index movement using Trend Deterministic Data Preparation and machine learning techniques," *Expert Syst. Appl.*, vol. 42, pp. 259-268, 2015.
- [17] J. Zhang, "Investment risk model based on intelligent fuzzy neural network and VaR," *J. Comput. Appl. Math.*, vol. 371, pp. 112707, 2020.
- [18] Y. Liang, D. Quan, F. Wang, X. Jia, M. Li, and T. Li, "Financial Big Data Analysis and Early Warning Platform: A Case Study," *IEEE Access*, vol. 8, pp. 36515-36526, 2020.
- [19] J. A. Lassa, A. Surjan, M. C. Anthony, and R. P. Fisher, "Measuring political will: An index of commitment to disaster and climate risk reduction," *Int. J. Disast. Risk Re.*, vol. 34, pp. 64-74, 2019.
- [20] J. Wang, and S. Xie, "Application of BP Neural Network in Early-Warning Analysis of Investment Financial Risk in Coastal Areas," *J. Coastal Res.*, vol. 106, pp. 259-262, 2020.
- [21] L. Zhu, M. Li, and N. Metawa, "Financial Risk Evaluation Z-Score Model for Intelligent IoT-based Enterprises," *Inform. Process. Manag.*, vol. 58, pp. 1-10, 2021.
- [22] S. A. Bondarenko, and V. V. Rummo, "Formation of the early warning system in management of risks of vine-fishing enterprises," *Econ. Innov.*, vol. 19, pp. 28-37, 2017.
- [23] W. Huang, Y. Yu, C. Tong, M. Xu, and R. Zhang, "Using a Duffing control approach to control the single risk factor in complex social-technical systems," *Inform. Sciences*, vol. 596, pp. 264-279, 2022.
- [24] X. Nie, and G. Deng, "Enterprise Financial Early Warning Based on Lasso Regression Screening Variables," *J. Financ. Risk Manag.*, vol. 09, pp. 454-461, 2020.
- [25] R. Dash, and P. K. Dash, "Efficient stock price prediction using a Self Evolving Recurrent Neuro-Fuzzy Inference System optimized through a Modified Differential Harmony Search Technique," *Expert Syst. Appl.*, vol. 52, pp. 75-90, 2016.

# Improving Privacy Preservation Approach for Healthcare Data using Frequency Distribution of Delicate Information

Ganesh Dagadu Puri, D. Haritha

Department of Computer Science and Engineering  
Koneru Lakshmaiah Education Foundation  
Vaddeswaram, AP, India

**Abstract**—In the modern world, everyone wishes that their personal information wouldn't be made public in any manner. In order to keep personal information hidden from prying eyes, privacy protection is essential. The data may be in the form of big data and minimization of risk and protection of sensitive data is important. In this research, a revolutionary customized privacy-preserving method is implemented that addresses the drawbacks of earlier personalized privacy as well as other anonymization methods. There are two main components that make up the proposed method's core. Delicate Information and Delicate Weight are two additional attributes which are used in the record table, are covered in the first section. The record holder's Delicate Information (DI) decides whether or not secrecy should be kept or if it should be shared. How delicate an attribute value is compared to the rest is indicated by its Delicate weight (DW). The second part covers a new representation used for anonymization termed the Frequency Distribution Block (FDB) and Quasi-Identifier Distribution Block (QIDB). According to experimental findings, the proposed system executes more quickly and with less data loss than current approaches.

**Keywords**—Privacy preservation approach; quasi identifier distribution block; frequency distribution block; big data; anonymization

## I. INTRODUCTION

Electronic Medical Records (EMRs) are currently widely used in healthcare networks. It makes it possible for people to easily and adaptably exchange their medical data. For instance, instead of needing to search through multiple physical records, a patient or his/her physician merely needs to access the data from a database to locate their diagnostic report. Advanced electronic medical record systems face a significant issue when it comes to securely storing and accessing electronic medical records because healthcare information is so sensitive [1]. Hadoop and big data analytics play a significant part in analyzing and processing the patient information in many forms to provide potential uses [2]. Investigation can leverage private data from several organizations to identify patterns. For instance, if a patient's private data is available across various hospitals, researchers can utilize it to better understand the patterns associated with a given disease and, as a result, make a more accurate diagnosis. The unprocessed information found in hospitals includes specific information on the patient, such as identity, address, date of birth, zip code, symptoms and

illness[3]. Before being delivered to the data receiver, the name and residential address information that are deemed private are stripped from the raw data which is also known as micro data. Furthermore, this micro data includes information like postal code and date of birth that can be connected to other external, publicly accessible data bases to re-identify sensitive value[4]. Linking attack refers to the process of re-identifying a record by connecting published data to publicly available data. Let us consider the patient records released by the hospital in Table I, for instance, which excludes data like name, residential data, and other private details. By joining the information from Table I with the publicly accessible external data base given in Table II, the intruder can disclose personal information. The query may appear like,

Select name, disorder from external\_table as A, patient\_table as B where A.postal=B.postal and A.age=B.age;

Since people are reluctant to volunteer their private data, it is extremely concerning that the answer to this query provides complete data about the illness and the name of the person. The join, which is referred to as Record Level Disclosure, may provide a value for age 36 and postal code 38677. Researchers employ techniques categorized as Privacy Preserving Data Publishing (PPDP) to hide confidential material from recipients. Quasi-Identifier (Q) attributes are characteristics found in Released Patient Data that can be connected to external, publicly accessible data bases, such as Postal Code, Date of Birth, etc. Data is modified in a way that leads to duplicate rows in the resulting table, limiting disclosure. Through the use of generalization, there has to be more than one implicit connection to the external data base. Thus, the k-anonymity algorithm is implemented for measuring this. Each entry in a table is indistinguishable from minimum k-1 other entries with regard to each and every set of quasi-identifier attributes if it fulfills the k-anonymity condition; such a table is known as a k-anonymous table.

With personalized anonymization, a guard node is utilized to determine if the record holder is willing to disclose the level of sensitivity upon which the anonymization will be carried out. As the record owner sensitivity is a generic one, the majority of the sensitive values that are included in the secret data base do not necessitate privacy protection. Therefore, just a small portion of the distribution's records need to be private. For instance, a record holder with malaria will not really mind



sharing his identity, in contrast to a record holder with HIV. The fact that some HIV-positive record owners are willing to expose their identities justifies this proposed privacy preservation strategy.

In other words, every group of quasi-identifier values needs a minimum  $k-1$  records, and they can be tricked by connecting a record from the disclosed data to a database with many entries that is publicly accessible. A two-anonymous generalization for Table I is shown in Table III. Assuming that the intruder uses a publicly accessible database and discovers that Ramesh is 36 years old with a postal code of 38677 and that he has a disorder, the intruder looks at anonymized Table III and learns that 38677 and 36 have been generalized to 386\*\* and [30-40] which can be associated with two entries of releases table and that the disorder cannot be derived from this information. Lung disease has been hidden and is not intended for publishing in this table ( $\langle 386^{**}, [50-60], \text{Lung Disease} \rangle$ ). Similar findings occur if the intruder attempts to infer Sitaram's illness, which belongs to category 3, but since every member of the category possesses the same sensitive property, the attacker deduces that Sitaram has fever.

Attribute level disclosure results from this release of confidential information. This occurs when a set of disorders are indeed symptoms of the same condition. To tackle this issue  $l$ -diversity was introduced. If the sensitive characteristic has at minimum 1 "well-represented" values, then an equivalence class has  $l$ -diversity. If each equivalence class in a table possesses  $l$ -diversity, the table has  $l$ -diversity. Additionally, skewness and similarity attacks are a drawback of  $l$ -diversity. Proximity was viewed as a method of overcoming this. The distribution of sensitive attributes in this strategy must match the anonymized chunk. Thus, there is a data loss.

In this paper, the research work is arranged into five sections. In Section 2, related work of various researchers and research limitations are described in detail. In Section 3, our proposed model is discussed. Experimental findings and discussion of each test is described in Section 4. Thus, in Section 5, research work is concluded and future scope of work is discussed.

TABLE I. PATIENT RELEASED DATA

Postal Code	Age	Disorder
38677	36	Mouth ulcer
38602	38	Brain cancer
38678	42	Fever
38685	46	Fever
38905	52	Fever
38906	56	Fever
38909	53	Fever
38673	58	Lungs Disease
38607	65	Lungs Disease
38655	68	Brain cancer

TABLE II. EXTERNAL DATABASE

Name	Postal Code	Age
Ramesh	38677	36
Laxmi	38677	45
Suresh	38602	38
Nagesh Rao	38602	32
Anupama	38678	42
Sitaram	38905	52
Kishor	38909	53
Vijay	38906	56
--	--	--

TABLE III. ANONYMOUS DATA

Postal Code	Age	Disease
386**	[30-40]	Mouth ulcer
386**	[30-40]	Brain cancer
386**	[40-50]	Fever
386**	[40-50]	Fever
389**	[50-60]	Fever
389**	[50-60]	Fever
389**	[50-60]	Fever
386**	[50-60]	Lungs Disease
386**	[60-70]	Lungs Disease
386**	[60-70]	Brain cancer

## II. LITERATURE SURVEY

Two anonymous techniques were presented by Xingguang Zhou et al. [5] that not only ensure data secrecy but also realize anonymity for patient. When attackers select attack destinations before gathering data from the electronic health record, the first strategy obtains modest security. The second strategy ensures total security by having attackers select attack targets in an adaptive manner upon contact with the electronic medical record system. It also suggested a method for EMR holders to use an anonymous search engine to find their electronic health records. As per Safa Bahri et al. [6] enormous amount of information, especially clinical data, has recently been amassed as a result of the intensification of emerging innovations that the large majority of people in the globe have accepted. Medical associations have acquired and analysed this clinical information and gain information and ideas that may be used to a variety of clinical judgments, including recommendations for medications and improved diagnoses. This paper mentions the significant effects that Big Data has on healthcare stakeholders, including patients, doctors, pharmaceutical and medical technicians, and medical insurance companies. It also examines the various difficulties that must be overcome in order to maximize the advantages of all the Big Data and the software that are presently accessible. Such large data can be stored on the devices customized to application processing [7].

A Secured as well as Anonymous Biometric Based User Authentication technique is introduced by B D Deebak et al. in 2017 [8] to guarantee secure data transmission in medical applications. This study demonstrates that a hostile cannot pretend to be a registered user in order to get unauthorized entry to or revoke an intelligent mobile card. For the purpose of demonstrating security and energy efficiency in healthcare application systems, a formal study relied on the random-oracle approach and resource evaluation is presented. The suggested method also incorporates some efficiency study to demonstrate that it offers high-security characteristics for developing intelligent medical application systems in the IoM. In 2019, Jorge Bernal Bernabe et al. [9] conducted a thorough evaluation of the State-of-Art (SoA) for privacy-preserving research approaches and methods in blockchain, and also the primary connected privacy issues in this exhilarating and disruptive technology. The survey includes privacy strategies in permissioned and privatized blockchains along with privacy-preserving research report and methods in accessible and private blockchain, such as Bitcoin and Ethereum. The analysis of various blockchain use cases includes looking at areas including Electronic-Government, Electronic-Health, crypto currency, developed cities, and cooperative ITS.

A Privacy-Preserving-Reinforcement-Learning (PPRL) architecture for the cloud computing system is proposed by Jaehyoung Park et al. in 2020 [10]. The proposed methodology makes use of learning with errors based cryptosystem for completely homomorphic encryption. Various cloud computing dependent intelligent service contexts are used to carry out effective analysis and assessment for the developed PPRL architecture. A stateless cloud monitoring approach for non-manager adaptive group data with preserving the privacy is proposed by Xiaodong Yang et al. [11]. With the random masking approach, the proposed methodology not only achieves individual identity privacy preservation but also data confidentiality preservation. Marwa Keshk et al. [12] present a thorough analysis of the most recent privacy-preserving methods for defending Cyber Physical System (CPS) technologies and their data against online threats in 2021. The ideas of privacy preservation and CPSs are examined, with an emphasis on the parts of cyber physical systems and how these systems might be hacked physically or digitally. Abdullah Al Omar et al. [13] presented an approach for the healthcare system which ensures data security and transparency. Additionally, the Ethereum platform is used to integrate insurance policies into the suggested system's blockchain, and cryptographic techniques are used to protect private information.

A mathematical formulation for an identity-based encryption strategy for the protection of patient confidentiality during the gathering of clinical records for evaluation is presented by Kissi Mireku Kingsford et al. in 2017[14]. The submission of medical data for analysis is becoming an essential element of daily life. To protect the confidentiality of patient, the model dissociates the identity of the patient from the investigated data upon data submission. A thorough analysis of privacy protection in big data from the communication point of view is presented by Tao Wang et al. in 2018[15]. It focuses on privacy-preserving methods,

especially differential privacy, and the basic privacy-preserving paradigm. Additionally, it examines the difficulties with differential privacy as well as its variations and modifications for various novel apps. Muneeb Ul Hassan et al. [16] have performed a detailed analysis of differential privacy approaches for CPSs as presented in 2019. Specifically, it looks at how differential privacy is used and implemented in four key CPS uses: energy, medical, transportation, healthcare & industrial Internet of things. It also outlines unresolved problems, difficulties, and prospective research directions for CPS differential privacy approaches. This investigation can be used as the foundation for the creation of cutting-edge differential privacy methods to handle numerous issues and CPSs' data privacy contexts.

The privacy of Kim's approach was assessed by Kefei Mao et al. [17], who show that the plan is actually vulnerable to the stolen smart - card threat. The plan also has some impassable stages, and the privacy assumption is excessively rigid. In addition, a novel technique built on Kim's as well as the quadratic residue hypothesis is investigated. In contrast to the current plans, the latest proposal does not call for the electric medical record database to personally communicate different secure values with patients and physicians. As a result, it is more useful and practical. It demonstrates that the suggested approach can offer greater protection than Kim's earlier plan. A unique architecture to enable privacy-preserving Machine learning (ML) was proposed by Kaihe Xu et al.[18], where the training data are spread and every shared data chunk is of enormous volume. To accomplish privacy preservation, it actually makes use of the Apache Hadoop platform's data locality attribute and just a few cryptographic functions at the Reduce functions. The comprehensive simulations used to show the presented strategy's robustness and consistency demonstrates that it is safe in the semi-honest framework.

In 2017, Tanashri Karle et al. [19] focused on protecting privacy by utilizing an anonymization methodology and a thorough investigation of two anonymization techniques are discussed namely - Datafly Technique and the Mondrian Algorithm. While Mondrian method is more suited for real datasets, Datafly technique is better suited for synthetic datasets. By using privacy preservation on a medical dataset in 2017, Balaji K. Bodkhe et al. [20] preserve a person's identity and any associated disorders (sensitive feature). The techniques including slicing, generalization, suppression and bucketization are utilized. These techniques guarantee privacy preservation while maintaining the usefulness of the data. The goal of S.Sathya et al. [21] is to take advantage of the new privacy difficulties posed by big data and focus on effective, privacy-preserving computation in the big data era. In order to address the effectiveness and privacy needs of Data Mining (DM) in the big data era, it first formalizes the overall framework of big data analytics, identifies the related privacy requirements, and introduces an effective and Privacy-Triple-DES as an instance.

To minimize and protect the data from unwanted parties, S. Shimona et al. [22] offer the PPDM strategies in a concise manner together with other privacy preservation measures in 2020. In 2020, Suneetha V et al. [23] introduced a unique concept called spark that uses Apache Spark to manage big data in the health care industry quickly and effectively while

using K-anonymization as well as L-diversity to disguise private data. The suggested method ensures that shared information will not reveal the actual information and that sensitive data is separated before being sent to Hadoop distributed file system. In 2019, Hui Jiang et al. [24] highlighted the fundamental steps of Hadoop-based big data analysis and included technical recommendations for common actual and off-line application scenarios. These recommendations were based on a review of the ecological structure of Hadoop. In order to have some reference value for the development of a big data platform and for the analysis and processing of huge data, Hadoop was utilized to construct the application context and the WordCount scenario was merged to assess the MapReduce calculation procedure.

A cooperation privacy preservation strategy for wearable technology was developed in 2018 by Hong Liu et al. [25] with id validation and data access control concerns in the space and time-aware settings. To obtain a secure healthcare pathway query under e-medical cloud servers without disclosing the secret data of patients like name, sex, age, location and also the information of hospitals like diagnosis, medication, and cost. Mingwu Zhang et al. [26] suggested a Privacy Preserving Enhancement of medical pathway query method. To maintain confidentiality in the e-Healthcare system, the suggested methodology first develops a number of privacy-preserving protocols like privacy-preserving medical comparison, privacy-preserving phase selection, and privacy-preserving phase update. It then implements the greedy approach in a secure way to carry out the query as well as the Min-Heap innovation to make it more efficient. This approach is feasible and effective with regard to computational time and cost, according to test findings. In 2018, Abdulatif Alabdulatif et al. [27] set out to propose a cloud-based solution for real-time patient monitoring that protects user privacy by spotting changes in a variety of important health indicators of participants of smart communities. IoT-enabled wearable devices' produced vital sign information is analysed in real-time on the cloud. The construction of a predictive method for the smart community while taking into account the sensitivity of information processing in a third-party context is the main topic of this paper (e.g., cloud computing). For enabling data prediction with patterns, it designed a crucial sign change detection method employing Holt's linear trend approach, where completely homomorphic encryption technique is applied to carry out calculations on an encrypted area that may protect data privacy. Additionally, a parallel strategy for encrypted operations using the MapReduce method of Apache Hadoop was proposed in order to minimize the burden of the completely homomorphic encryption technique across massive healthcare data.

The difficulties and needs of creating frameworks and procedures for globally distributed data processing are investigated and discussed by Shlomi Dolev et al. in 2017 [28]. It categorizes and studies the overhead problems associated with batch, stream and SQL-style processing using geo-distributed architectures, methods, and techniques. Using differential privacy, Miao Du et al. [29] present and put into practice a ML technique for smart edges in 2018. In a wireless big data situation, anonymization in training datasets is the

main priority. Additionally, it designs two distinct techniques, Output and Objective Perturbation which fulfill differential privacy, and guarantees privacy and security by including Laplace techniques. Additionally, for correlated datasets, differential privacy preservation algorithms are offered, providing privacy through theoretical inference. Ultimately, tests were conducted using TensorFlow and the effectiveness of the technique was assessed using the four datasets STL-10, SVHN, MNIST and CIFAR-10. The suggested approach effectively ensures accuracy upon benchmark datasets while safeguarding the confidentiality of training datasets.

A scalable approach to the local-recoding issue for big data anonymization over proximity privacy violations was investigated by Xuyun Zhang et al. [30]. The study proposes a proximity privacy framework that provides the semantic proximity of sensitive values including numerous sensitive attributes. It also models the local recoding issue as a proximity-aware clustering issue. It presents a scalable two-phase clustering method that combines the proximity-aware agglomerative clustering technique and the t-ancestors clustering technique. The methods were created using MapReduce to provide good scalability using cloud-based data-parallel processing. Numerous tests using real data sets show that the method greatly outperforms existing methods in terms of scalability, time efficiency, and capacity to fight against proximity information leakage.

As per Haiping Huang et al. [31], Electronic-healthcare has substantially benefited from the industrialization of cloud computing, Internet of things and Wireless-body-Area-Networks (WBANs). Furthermore, there are still several obstacles standing in the way of e-Healthcare's growth, especially issues with data security and privacy protection. Healthcare system architecture is formulated to overcome these issues. It gathers health information from WBANs, transfers it across a substantial wireless sensor network, and then releases it into Wireless-Personal-Area-Networks (WPANs) through a gateway. Additionally, healthcare system uses the Homomorphic Encryption Dependent on Matrix scheme to assure confidentiality, the Groups of Send-Receive Model strategy to accomplish key distribution, and an intelligent system capable of autonomously analyzing the encrypted health data and reporting the findings. The confidentiality, privacy, and improved efficiency of healthcare system are evaluated theoretically and experimentally in comparison to existing systems or techniques. Lastly, the practicality of the healthcare system prototype implementation is examined. A privacy-preserving approach is put forth by Marwa Keshk et al. in 2019 [32] in order to obtain both safety and confidentiality in intelligent power networks. A two-level privacy component and an anomaly detection component are the framework's two core components. Using open datasets, the outlier detection module trains and validates the outcomes of the two-level privacy component using a Long-Short-Term-Memory DL approach. In contrast to various cutting-edge methodologies, the experiments demonstrated that the proposed architecture can effectively secure data of intelligent power networks and identify anomalous behaviors. The term "optimal distributed estimate" refers to a conceptual framework created by Jianping He et al. in 2018 [33] to examine how to maximize the

assessment of a neighbor's original data using the collected local data. The disclosure probability is then looked into as part of the best estimation for the data privacy evaluation. The privacy-preserving average consensus method's data privacy has been further examined using the established framework, and the best noises for the technique are identified.

In 2018, Weichao Gao et al. [34] used the idea of homomorphic encryption as well as secured network protocol development to tackle the issues of privacy preservation for information auction in CPS. A general Privacy-Preserving Auction Strategy is put forth, in which an unreliable third-party trade platform is made up of the two distinct entities of the auctioneer and interim platform. A winner in the auction procedure is defined and all bidder data is hidden by using homomorphic encryption as well as a one-time pad. However, it also suggests an Enhanced Privacy Preserving Auction Method that makes use of an extra signature verification technique in order to increase the overall security of the privacy preserving auction. Each strategy's viability is confirmed through in-depth theoretical analysis and thorough performance tests, which also include an examination of attack tolerance. A unique privacy-preserving anomaly - based detection methodology, known as PPAD-CPS, is suggested by Marwa Keshk et al. in 2018 [35] for safeguarding private data and identifying hostile findings in power technology and associated network traffic. There are two primary components in the architecture. In order to meet the goal of privacy preservation, a data pre-processing component is first proposed for filtering and changing original information into a new format. Secondly, an anomaly-based detection component utilizing a Kalman Filter as well as Gaussian Mixture Model for accurately predicting the posterior probabilities of normal and malicious events is proposed. Two open datasets, the Power System as well as UNSW-NB15 dataset, are used to test the efficiency of the architecture.

### III. PROPOSED SYSTEM

The privacy-preserving method we propose overcomes the drawbacks of existing techniques and other anonymization methods. There are two main parts that make up the proposed method's core. The first part of the equation concerns with additional attributes utilized in the table namely Delicate Information and Delicate weight. The DI indicates whether the privacy of the record owner's private data should be protected or released. DW determines the sensitivity of the attribute. DW is necessary for DI.

When the person provides their data, DI can be accessed easily from them. DW could be based on previously acquired sensitive attribute information. The same level of protection is provided for every sensitive attributes by conventional privacy approaches, which has been addressed in this approach by the implementation of DI and DW. The flag  $DI=0$  indicates that the entry holder is not willing to share his confidential attribute, while  $DI=1$  indicates that he has no problem doing so. The publisher has highlighted DW for any sensitive attributes where confidentiality is crucial. For instance, a record holder with the fever or gastroenteritis is less reluctant to expose his identify than a label owner with cancer. Whenever the sensitive attribute is a very common disorder

like the fever or mouth ulcer,  $DW=0$  is being used; for a sensitive attribute like brain cancer, which is uncommon,  $DW=1$  is utilized. For  $DW=0$ , DI has a default value of 1, and for  $DW=1$ , the record holder's DI values are accepted.

TABLE IV. DW FOR DISORDERS

Disease	DW
Mouth ulcer	0
Brain cancer	1
Fever	0
Lungs Disease	1

The second section discusses a novel approach for evaluating the distribution known as the FDB and QIDB. Each disorder's spread in the FDB is based on original, personal data. QIDB is formed for each entry with  $DW=1$  and  $DI=0$ . Several QIDB chunks will exist. These chunks are needed to make sure that each particular QIDB and distribution of FDB is synchronized.

TABLE V. PATIENT RELEASED DATA WITH DW AND DI

Postal Code	Age	Disorder	DW	DI
38677	36	Mouth ulcer	0	1
38602	38	Brain cancer	1	0
38678	42	Fever	0	1
38685	46	Fever	0	1
38905	52	Fever	0	1
38906	56	Fever	0	1
38909	53	Fever	0	1
38673	58	Lungs Disease	1	1
38607	65	Lungs Disease	1	0
38655	68	Brain cancer	1	1

TABLE VI. FREQUENCY DISTRIBUTION BLOCK

Disease	Probability
Mouth ulcer	0.1
Brain cancer	0.2
Fever	0.5
Lungs Disease	0.2

#### A. Model and Terminology for Proposed Personalized Privacy

Let R be a connection providing personal information about a set of people. There are four groups of attributes in R.

- Unique Identifiers  $U_j$  - It can be used to identify individuals who are eliminated from R.
- Quasi identifiers  $QI_j$  - its value can be combined with publicly available information to determine a person's identity.

- Delicate attributes  $D_j$  – It is secretive or delicate to the record holder.
- Non quasi identifiers  $NQI_j$  – It doesn't fall into any of the three categories.

The goal of proposed method is to obtain a generalized table  $R^*$  such that distribution of every QIDB is comparable to the diversity of the entire distribution as seen in FDB. For ease of use, the full set of quasi identifiers is denoted by  $QI$ , and its values by  $q$ . In a similar manner, there is a single delicate attribute  $D_i$  and its value  $d$ . Relation  $R$  comprises of  $m$  number of tuples  $R = \{r_1, r_2, \dots, r_m\}$ . Record holder data can be obtained by referring as  $r_j.d$  to represent delicate value and  $r_j.q$  for quasi identifier value  $1 \leq j \leq m$ .

1) *Delicate Weight*- for every tuple  $r \in R$ , its delicate weight is added. This value is derived from Relation  $W(ds, dw)$  where  $ds$  indicates disorder and  $dw$  indicates delicate weight.  $W$  contains  $p$  records

$$r_j.dw = \{ w_j.dw \text{ if } w_j.ds = r_i.d \mid 1 \leq i \leq p \} \text{ for every } 1 \leq j \leq m$$

Table IV provides the  $dw$  value for every disorder. Table I is used to create this distribution.

2) *Delicate Information* - for every tuple  $r \in R$ , its Delicate Information is indicated as  $r.di$ .

$$r_j.di = \{ 1 \text{ if } r_i.dw = 0 \text{ and } r_j.dw = 1 \} \text{ for every } 1 \leq j \leq m$$

The value of user defined ( $ud$ ), is either 0 or 1. If the value of  $r_i.di$  is zero, the user is not prepared to share his information, and if it is one, the user agrees.

Table V shows the values of  $dw$  and  $di$  assuming that the record holder will approve  $di$  value for  $DW=1$ . Additionally, it can be seen that if  $dw=0$ , the corresponding  $di$  is set to 1, showing that the entries' sensitivity is not really important.

3) *Thresholds* - To improve and enhance effectiveness of disclosure, generalization, and suppression, values of threshold are established for a number of personalized privacy aspects.

- $T_n$  - It indicates minimum number of entries in  $R$ .
- $T_{itr}$  - It indicates maximum number of required iterations.
- $T_{sup}$  - It indicates minimum number of delicate values for suppression.
- $T_{dis}$  - It indicates minimum number of delicate values for disclosure
- $T_{acc}$  - It indicates minimum number of thresholds for addition or subtraction.

Several threshold values are suggested because the dispersion aspect is being taken into account. The first value, which was never specified in the earlier representations, denotes the bare minimum number of item sets that must be provided in order to execute anonymization.  $T_{itr}$  is calculated using information of the Value domain hierarchy's height. The generalization is greater and information loss is correspondingly greater when the value of  $T_{itr}$  is high.  $T_{sup}$

denotes the absolute minimal amount of sensitive distribution that could exist in QIDB for that block's deletion following  $T_{itr}$ . The threshold value  $T_{dis}$  represents the amount that can be added or removed from every frequency distribution for every disorder in order to make it equal to the FDB distribution. The frequency of QIDB and FDB will not be completely the same, thus while examining the distribution of every disorder is examined if the frequency in that  $qidb.v.d \pm T_{acc}$  always  $T_{dis} > T_{acc}$ .

4) *Frequency Distribution Block* - Distribution of every  $w_j.ds$  in regards to the original distribution  $r_i.d$  is stored in relation  $FDB(ds, p)$  where  $d$  represents disorder and  $p$  represents probability distribution of it Every  $p$  for  $ds$  is computed by mapping every  $ds$  in  $R$  (values of  $r_i.d = fdbv.ds$ ) to the total no. of tuples in  $R$ , for every  $1 \leq v \leq k$ . Considering there are  $m$  entries in the relation.

5) *Quasi - Identifier Distribution Block*- for every  $r_j.d$  where  $r_j.dw=1$  &  $r_i.di=0$  a new QIDB is generated comprising  $r_i.s$  for every  $1 \leq j \leq m$ . The relation  $QIDB.V(q, d)$  where  $qidb.vl.q = r_j.q$  &  $qidb.vl.d = r_j.d$ . Considering there are  $m$  QIDB chunks.

TABLE VII. QIDB.1 DATA

Postal Code	Age	Disorder
38602	38	Brain cancer

TABLE VIII. QIDB.2 DATA

Postal Code	Age	Disorder
38607	65	Lungs Disease

Table VI illustrates the frequency distribution of every disorder. This distribution demonstrates that the fever is a widespread disorder with a higher frequency—roughly 50 percent in the reported data. Every QIDB maintains the exact similar distribution. Due to the fact that the quasi values  $\langle 38602, 28 \rangle$  and  $\langle 38607, 55 \rangle$  have the  $DW$  and  $DI$  values of 1 and 0 respectively, in the first cycle 2 blocks of QIDB will be produced for these values as shown in Table VII and Table VIII. Table VII shows Brain cancer disease probability is 0.2 in distribution block. In the same way Table VIII shows Probability of Lungs disease is 0.2 in the frequency distribution block. It is calculated from delicate weight of delicate information.

6) *Generalization* - A generalization function provides the general domain of an attribute  $R.Q$ . Function will return a generalized value in the domain provided a value  $r.q$  in the original domain.

7) *Check Frequency*- for every QIDB, examine  $CF_q(QIDB.V)$  with  $QIDB.V$  FD which is equal to the FD in FDB. It is performed as follows

Let  $c$  be the total number of entries in  $QIDB.V$  for every  $UNI_q(qidb.vl.d)$  obtain total number of mappings which match  $qidb.vl.d$  to the total number of entries that is  $x$  in  $QIDB.V$ , thus  $CF_q$  will return true if

For every  $1 \leq v \leq m$  such that  $fdbv.ds=qldb.vl.d$

$$fdbv.p = (\text{unique}(qldb.vl.d) / x) \pm Tacc$$

This is examined in each cycle if a QIDB satisfies the FD then this chunk won't be taken into account for the next iteration.

8) *Suppression*- After  $Titr$  iterations,  $SUP(QIDB.v)$  remove the chunk if it meets the following criteria

For every  $1 \leq u \leq m$  such that for every  $fdbu.ds=qldb.vl.d \wedge fdbu.ds=wj.ds \wedge wj.dw=1$  for every  $j \ 1 \leq j \leq k$

$$\text{Count}(qldb.vl.d) \leq Tsup$$

9) *Disclosure* - After  $Titr$  iterations,  $DIS(QIDB.v)$  adds extra records if it meets the following criteria for every  $1 \leq l \leq x$  such that for every  $fdbu.ds=qids.vl.d \wedge fdbu.ds=wi.d \wedge wi.dw=1$  for some  $i \ 1 \leq i \leq k$

$$(\text{unique}(qldb.vl.d) / x) = Tdis \pm fdbv.p$$

### B. Personalized Privacy Breach

Assume an attacker who tries to estimate important information from a record holder  $h$ . In the worst situation, the attacker only pays attention to the tuples  $r^* \in R^*$  whose  $Q$  value  $rj^*.q$  covers  $x.q$  for all  $j$  such that  $1 \leq j \leq n$  since it is assumed that the adversary knows  $Q$  of  $H$ .  $Q$ -group is formed by these tuples. That is, if  $rj^*$  and  $rjp^*$  are two such tuples then  $rj^*.q=rjp^*.q$  for all  $j$  such that  $1 \leq j \leq n$ . The adversary cannot deduce a sensitive attribute of  $h$  if this group is not established.

1) *Required Q-Group/ Actual (h)* - Given an individual  $h$ , the Required  $Q$ -group  $ReqG(H)$  is the only  $Q$ -group in  $r^*$  covers  $h.q$ . Considering  $Actual(X)$  represents those records which are generalized to  $RG(H)$ .

The attacker has no knowledge about  $Actual(H)$ . To acquire  $Actual(H)$ , the adversary must locate some external data base  $External(H)$  that should be covered in  $ReqG(H)$ .

2) *External DataBase Ext (x)*-  $External(H)$  is a collection of individuals whose value is covered by  $ReqG(H)$ .

$$Actual(H) \subseteq External(H)$$

The adversary uses a combinational strategy to deduce sensitive attribute of  $h$ . let us consider that  $h.s$  is present in one of  $ri^*$  and  $h$  is not repeated. The possible reconstruction of the  $ReqG(X)$  contains  $h$  different record holders  $h1, h2, h3, \dots, hr$  who belong to  $External(H)$  but there can be only  $y$  in  $ReqG(H)$ . This can be seen by the probabilistic nature and can be represented as  $perm(x,y)$ .

$perm(x,y)$  is Possible Reconstruction that can be created by with  $h$  holders and  $y$  mappings. Breach Probability represents the probability of inferred information. Let us consider  $Actual N$  represents actual number of entries with sensitive attribute from which  $h$  can be deduced.

$$\text{Breach probability} = \text{Actual } N / perm(x,y)$$

Breach probability will decide the privacy factors, If it is 100 percent then  $h$  can be deduced; if it is poor then the inference will be tough for the adversary.

### C. Quasi-Identifier Distribution Block - Anonymization Algorithm

Since it is assumed that the sensitivity distribution in every location is typically fairly uniform, this technique processes quasi values sequentially. Consider the following algorithm of QIDB.

---

#### Algorithm 1: QIDB-Anonymization

---

**Input:** personal data  $R$  with DW-DI, threshold values  $T_n, T_{itr}, T_{sup}, T_{dis}, T_{acc}$  and initialized  $FDB(ds,p)$

**Output:** Released table  $T^*$

**Step 1:** if  $(n < T_n)$  then return value 1

**Step 2:** for each  $rj.s$  where  $rj.dw=1$  &  $rj.di=0$  a new QIDB is generated comprising  $rj.d$  and  $rj.q$  for every  $1 \leq j \leq n$ .

**Step 3:**  $initial\_iteration=0,$   
 $receive\_flag=0$   
 $gen=Initial G(R)$

**Step 4:** while  $(initial\_iteration < T_{itr}$  and  $receive\_flag=0)$   
QIDB chunks are deleted if  $CFq()$  returns true then examines the value of QIDB if it is 0 then  $receive\_flag=1$   
 $Iteration = iteration + 1$   
 $gen = next G(R)$

**Step 5:** if  $receive\_flag=0$  then  
execute  $sup()$  and  $dis()$

**Step 6:** Examine value of QIDB if it is 0 then  $receive\_flag=1$

**Step 7:** release  $R^*$  if  $receive\_flag=1$

---

The resultant anonymization after implementing Personal Anonymization of one of the QIDB with  $T_{acc}=0.1$  chunk is depicted in Table IX.

TABLE IX. RESULTANT DW-DI BASED QIDB ANONYMIZATION WITH  $T_{acc}=0.1$

ZIP Code	Age	Disorder
386**	[30-50]	Brain cancer
386**	[30-50]	Mouth ulcer
386**	[30-50]	Lungs Disease
386**	[30-50]	Fever
386**	[30-50]	Fever

## IV. EXPERIMENTAL FINDINGS AND DISCUSSION

Effectiveness of proposed method in comparison to  $k$ -anonymity as well as  $l$ -diversity is obtained. The investigation made use of a common dataset. 400-records of adult dataset are taken into account with the relevant quasi-attributes: age, gender, marital status, and profession. Age is the only attribute that is numerical; all other attributes are categorical. For  $DW=1$ , probability is utilized to determine the DI value.

In Fig. 1, it is shown that data loss for proposed method is less than  $k$ -anonymity and  $l$ -diversity. Number of records can be increased in the proportion to see the information loss in three methods and compare it.

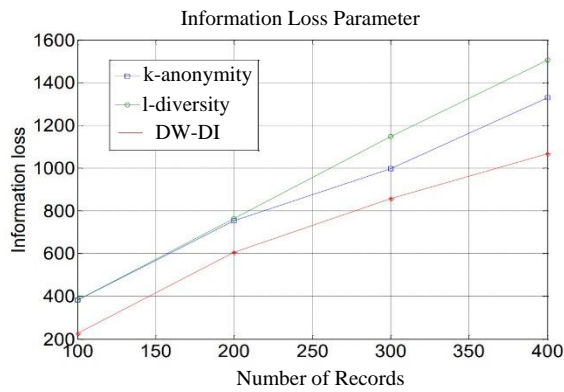


Fig. 1. Information Loss of DW-DI Proposed Personal Anonymization Technique Compared with l-Diversity and k-Anonymity Technique.

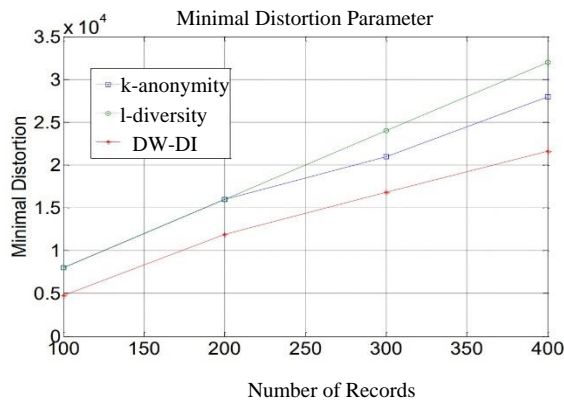


Fig. 2. Minimal Distortion Parameter of DW-DI Personal Anonymization Compared to l-Diversity k-Anonymity Technique.

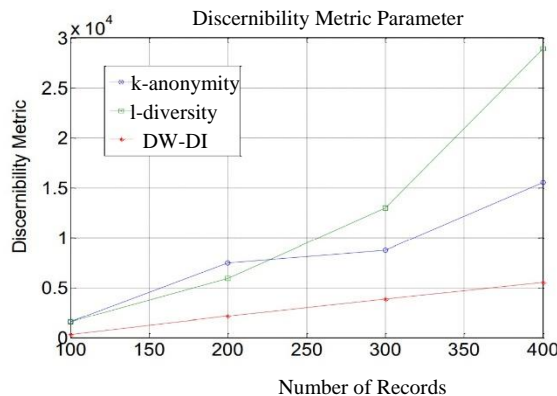


Fig. 3. Discernibility Metric Parameter of DW-DI Personal Anonymization Compare to l-Diversity and k-Anonymity Technique.

For quasi identification, a generalization hierarchy is created and employed and a distance vector is produced and is used in this approach. The generalization hierarchy can go up to a maximum level of 10. In Fig. 1, the information loss factor is displayed. The data quality improves when there is less data loss. The concept of minimal distortion centers on penalizing every value that has been generalized or repressed. When a hierarchy inside the domain generalization hierarchy is extended to the next level, it is given a penalty. In Fig. 2,

minimum distortion is displayed. A penalty of 10 is applied in test for each generalization. Fig. 3 illustrates how this Discernibility Metric determines the cost by penalizing every tuple for being unrecognizable from other tuples. In Fig. 4, runtime is displayed. For the test, the threshold values  $T_n = 400$ ,  $T_{itr} = 10$ ,  $T_{dis} = 0.01$ ,  $T_{sup} = 1$ ,  $T_{acc} = 01$  was used.

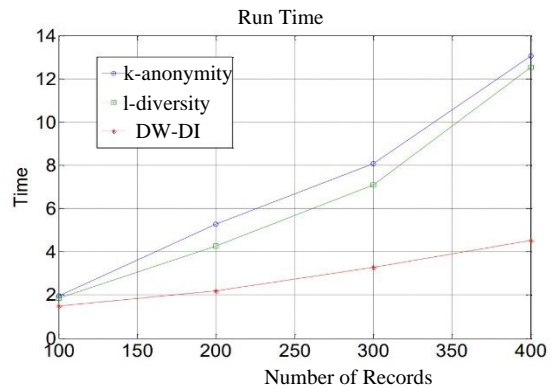


Fig. 4. Run Time of DW-DI Personal Anonymization Compare to l-Diversity and k-Anonymity Technique.

## V. CONCLUSION AND FUTURE WORK

Since the runtime and quality of the data are better with personalized privacy, it is an essential research direction. Because all entries do not need to be private, using DW not only enhances the signal of sensitivity but also increases the usefulness of the data. Since many of the record holders are willing to expose their identities, DI is an extra flag that increases the quality of the data in the DW record. Therefore, DW-DI is a better solution for personalized privacy than employing a guarding node alone. Using anonymization depending on QIDB, several quasi groups can be separately generalized. This method improves confidentiality by checking each QIDB chunk for a FD of sensitive values that is roughly equivalent to the FD of sensitive values in the original contents. Additionally, it defeats probabilistic assault, attribute connection and record connection. When a specific sensitivity's frequency distribution is localized in a small area of an individual pattern, this method performs effectively.

Future research can go in a number of different ways as it examines QIDB anonymization of DW-DI personal privacy. Firstly, the impact of sequential and multiple distributions of released data have not been taken into account. Research on sensitivity weighting can be taken into consideration. In this method, records are processed sequentially to see if the generalized record fits the QIDB generalized value, and if they do, the record is added to the block. Different techniques can be investigated as an option to sequential processing. Multi-dimensional data and unorganized schema can both be used with this technique.

## REFERENCES

- [1] S. Kim, M. K. Sung, and Y. D. Chung, "A framework to preserve the privacy of electronic health data streams," *J. Biomed. Inform.*, vol. 50, pp. 95–106, 2014, doi: 10.1016/j.jbi.2014.03.015.
- [2] G. D. Puri and D. Haritha, "Survey big data analytics, applications and privacy concerns," *Indian J. Sci. Technol.*, vol. 9, no. 17, 2016, doi: 10.17485/ijst/2016/v9i17/93028.



- [3] G. D. Puri and D. Haritha, "Framework to avoid similarity attack in big streaming data," *Int. J. Electr. Comput. Eng.*, vol. 8, no. 5, 2018, doi: 10.11591/ijece.v8i5.pp.2920-2925.
- [4] G. D. Puri and D. Haritha, "A novel method for privacy preservation of health data stream," *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 9, no. 4, pp. 4959–4963, 2020, doi: 10.30534/ijatcse/2020/110942020.
- [5] X. Zhou, J. Liu, Q. Wu, and Z. Zhang, "Privacy Preservation for Outsourced Medical Data with Flexible Access Control," *IEEE Access*, vol. 6, pp. 14827–14841, 2018, doi: 10.1109/ACCESS.2018.2810243.
- [6] S. Bahri, N. Zoghliani, M. Abed, and J. M. R. S. Tavares, "BIG DATA for Healthcare: A Survey," *IEEE Access*, vol. 7, pp. 7397–7408, 2019, doi: 10.1109/ACCESS.2018.2889180.
- [7] P. Ganesh D, P. Dinesh D, and W. Manoj A., "RAID 5 Installation on Linux and Creating File System," *Int. J. Comput. Appl.*, vol. 85, no. 5, pp. 43–46, 2014, doi: 10.5120/14841-3107.
- [8] B. D. Deebak, F. Al-Turjman, M. Aloqaily, and O. Alfandi, "An authentic-based privacy preservation protocol for smart e-healthcare systems in iot," *IEEE Access*, vol. 7, pp. 135632–135649, 2019, doi: 10.1109/ACCESS.2019.2941575.
- [9] J. Bernal Bernabe, J. L. Canovas, J. L. Hernandez-Ramos, R. Torres Moreno, and A. Skarmeta, "Privacy-Preserving Solutions for Blockchain: Review and Challenges," *IEEE Access*, vol. 7, pp. 164908–164940, 2019, doi: 10.1109/ACCESS.2019.2950872.
- [10] J. Park, D. S. Kim, and H. Lim, "Privacy-preserving reinforcement learning using homomorphic encryption in cloud computing infrastructures," *IEEE Access*, vol. 8, 2020, doi: 10.1109/ACCESS.2020.3036899.
- [11] X. Yang, M. Wang, X. Wang, G. Chen, and C. Wang, "Stateless Cloud Auditing Scheme for Non-Manager Dynamic Group Data with Privacy Preservation," *IEEE Access*, vol. 8, pp. 212888–212903, 2020, doi: 10.1109/ACCESS.2020.3039981.
- [12] M. Keshk, B. Turnbull, E. Sitnikova, D. Vatsalan, and N. Moustafa, "Privacy-Preserving Schemes for Safeguarding Heterogeneous Data Sources in Cyber-Physical Systems," *IEEE Access*, vol. 9, pp. 55077–55097, 2021, doi: 10.1109/ACCESS.2021.3069737.
- [13] A. Al Omar et al., "A Transparent and Privacy-Preserving Healthcare Platform with Novel Smart Contract for Smart Cities," *IEEE Access*, vol. 9, pp. 90738–90749, 2021, doi: 10.1109/ACCESS.2021.3089601.
- [14] M. D. N. Ayeh, and A. M. Kissi Mireku Kingsford, Fengli Zhang, "A Mathematical Model for a Hybrid System Framework for Privacy Preservation of Patient Health Records," in 2017 IEEE 41st Annual Computer Software and Applications Conference (COMPSAC), 2017, pp. 119–124, doi: 10.1109/COMPSAC.2017.21.
- [15] S. Yao, and Z. Huo, T. Wang, Z. Zheng, M. H. Rehmani, "Privacy Preservation in Big Data From the Communication Perspective—A Survey," *IEEE Commun. Surv. Tutorials*, vol. 21, no. 1, pp. 753–778, 2019, doi: 10.1109/COMST.2018.2865107.
- [16] M. U. Hassan, M. H. Rehmani, and J. Chen, "Privacy preservation in blockchain based IoT systems: Integration issues, prospects, challenges, and future research directions," *Futur. Gener. Comput. Syst.*, vol. 97, 2019, doi: 10.1016/j.future.2019.02.060.
- [17] J. Liu, and M. Wang, Kefei Mao, Jie Chen, "Security enhancement on an authentication scheme for privacy preservation in Ubiquitous Healthcare System," in 2015 4th International Conference on Computer Science and Network Technology (ICCSNT), 2015, pp. 885–892, doi: 10.1109/ICCSNT.2015.7490882.
- [18] K. Xu, H. Yue, L. Guo, Y. Guo, and Y. Fang, "Privacy-Preserving Machine Learning Algorithms for Big Data Systems," *Proc. - Int. Conf. Distrib. Comput. Syst.*, vol. 2015-July, pp. 318–327, 2015, doi: 10.1109/ICDCS.2015.40.
- [19] T. Karle and D. Vora, "Privacy preservation in big data using anonymization techniques," 2017 Int. Conf. Data Manag. Anal. Innov. ICDMAI 2017, pp. 340–343, 2017, doi: 10.1109/ICDMAI.2017.8073538.
- [20] B. K. Bodkhe, "Privacy Preservation for Medical Dataset using Hadoop," in 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS), 2017, pp. 3463–3468.
- [21] S. Sathya and T. Sethukarasi, "Efficient privacy preservation technique for healthcare records using big data," 2016 Int. Conf. Inf. Commun. Embed. Syst. ICICES 2016, no. Icices, 2016, doi: 10.1109/ICICES.2016.7518878.
- [22] S. Shimona, "Survey on Privacy Preservation Technique," *Proc. 5th Int. Conf. Inven. Comput. Technol. ICICT 2020*, pp. 64–68, 2020, doi: 10.1109/ICICT48043.2020.9112584.
- [23] P. S. V., "Privacy Preservation of Healthcare Big Data," no. Icimia, pp. 743–749, 2020.
- [24] H. Jiang, "Research and practice of big data analysis process based on Hadoop framework," *Proc. 2019 IEEE 3rd Inf. Technol. Networking, Electron. Autom. Control Conf. ITNEC 2019*, no. It nec, pp. 2044–2047, 2019, doi: 10.1109/ITNEC.2019.8729522.
- [25] H. Liu, X. Yao, T. Yang, and H. Ning, "Cooperative privacy preservation for wearable devices in hybrid computing-based smart health," *IEEE Internet Things J.*, vol. 6, no. 2, pp. 1352–1362, 2019, doi: 10.1109/JIOT.2018.2843561.
- [26] M. Zhang, Y. Chen, and W. Susilo, "PPO-CPQ: A Privacy-Preserving Optimization of Clinical Pathway Query for E-Healthcare Systems," *IEEE Internet Things J.*, vol. 7, no. 10, pp. 10660–10672, 2020, doi: 10.1109/JIOT.2020.3007518.
- [27] A. Alabdulatif, I. Khalil, A. R. M. Forkan, and M. Atiquzzaman, "Real-Time Secure Health Surveillance for Smarter Health Communities," *IEEE Commun. Mag.*, vol. 57, no. 1, pp. 122–129, 2019, doi: 10.1109/MCOM.2017.1700547.
- [28] S. Dolev, P. Florissi, E. Gudes, S. Sharma, and I. Singer, "A Survey on Geographically Distributed Big-Data Processing Using MapReduce," *IEEE Trans. Big Data*, vol. 5, no. 1, pp. 60–80, 2017, doi: 10.1109/tbdata.2017.2723473.
- [29] M. Du, K. Wang, Z. Xia, and Y. Zhang, "Differential Privacy Preserving of Training Model in Wireless Big Data with Edge Computing," *IEEE Trans. Big Data*, vol. 6, no. 2, pp. 283–295, 2018, doi: 10.1109/tbdata.2018.2829886.
- [30] X. Zhang et al., "Proximity-aware local-recoding anonymization with MapReduce for scalable big data privacy preservation in cloud," *IEEE Trans. Comput.*, vol. 64, no. 8, pp. 2293–2307, 2015, doi: 10.1109/TC.2014.2360516.
- [31] H. Huang, T. Gong, N. Ye, R. Wang, and Y. Dou, "Private and Secured Medical Data Transmission and Analysis for Wireless Sensing Healthcare System," *IEEE Trans. Ind. Informatics*, vol. 13, no. 3, pp. 1227–1237, 2017, doi: 10.1109/TII.2017.2687618.
- [32] M. Keshk, B. Turnbull, N. Moustafa, D. Vatsalan, and K. K. R. Choo, "A Privacy-Preserving-Framework-Based Blockchain and Deep Learning for Protecting Smart Power Networks," *IEEE Trans. Ind. Informatics*, vol. 16, no. 8, pp. 5110–5118, 2020, doi: 10.1109/TII.2019.2957140.
- [33] J. He, L. Cai, and X. Guan, "Preserving data-privacy with added noises: Optimal estimation and privacy analysis," *IEEE Trans. Inf. Theory*, vol. 64, no. 8, pp. 5677–5690, 2018, doi: 10.1109/TIT.2018.2842221.
- [34] W. Gao, W. Yu, F. Liang, W. G. Hatcher, and C. Lu, "Privacy-Preserving Auction for Big Data Trading Using Homomorphic Encryption," *IEEE Trans. Netw. Sci. Eng.*, vol. 7, no. 2, pp. 776–791, 2020, doi: 10.1109/TNSE.2018.2846736.
- [35] M. Keshk, E. Sitnikova, N. Moustafa, J. Hu, and I. Khalil, "An Integrated Framework for Privacy-Preserving Based Anomaly Detection for Cyber-Physical Systems," *IEEE Trans. Sustain. Comput.*, vol. 6, no. 1, pp. 66–79, 2019, doi: 10.1109/tsusc.2019.2906657.

# Attention-based Long Short Term Memory Model for DNA Damage Prediction in Mammalian Cells

Mohammad A. Alsharaiah<sup>1\*</sup>, Laith H. Baniata<sup>2\*</sup>, Omar Adwan<sup>3</sup>, Ahmad Adel Abu-Shareha<sup>4</sup>, Mosleh Abu Alhaj<sup>5</sup>,  
Qasem Kharm<sup>6</sup>, Abdelrahman Hussein<sup>7</sup>, Orieb Abualghanam<sup>8</sup>, Nabeel Alassaf<sup>9</sup> and Mohammad Baniata<sup>10</sup>  
Al-Ahliyya Amman University; Amman, Jordan<sup>1, 3, 4, 5, 6, 7, 8</sup>  
Gachon University; South Korea<sup>2</sup>  
The University of Jordan<sup>9</sup>  
Ubion, South Korea<sup>10</sup>

**Abstract**—The understanding of DNA damage intensity – concentration-level is critical for biological and biomedical research, such as cellular homeostasis, tumor suppression, immunity, and gametogenesis. Therefore, recognizing and quantifying DNA damage intensity levels is a substantial issue, which requires further robust and effective approaches. DNA damage has several intensity levels. These levels of DNA damage in malignant cells and in other unhealthy cells are significant in the assessment of lesion stages located in normal cells. There is a need to get more insight from the available biological data to predict, explore and classify DNA damage intensity levels. Herein, the development process relied on the available biological dataset related to DNA damage signaling pathways, which plays a crucial role in DNA damage in the mammalian cell system. The biological dataset that was used in the proposed model consists of 15000 records intensity – concentration-level for a set of five proteins which regulate DNA damage. This research paper proposes an innovative deep learning model, which consists of an attention-based long short term-memory (AT-LSTM) model for DNA damage multi class predictions. The proposed model splits the prediction procedure into dual stages. For the first stage, we adopt the related feature sequences which are inserted as input to the LSTM neural network. In the next stage, the attention feature is applied efficiently to adopt the related feature sequences which are inserted as input to the softmax layer for prediction in the following frame. Our developed framework not only solves the long-term dependence problem of prediction effectively, but also enhances the interpretability of the prediction methods that was established on the neural network. We conducted a novel proposed model on big and complex biological datasets to perform prediction and multi classification tasks. Indeed, the (AT-LSTM) model has the ability to predict and classify the DNA damage in several classes: No-Damage, Low-damage, Medium-damage, High-damage, and Excess-damage. The experimental results show that our framework for DNA damage intensity level can be considered as state of the art for the biological DNA damage prediction domain.

**Keywords**—Mammalian cell; deep learning techniques; attention; LSTM; classification; DNA damage

## I. INTRODUCTION

Mammalian cells have a complicated organism system. Specifically, each cell has a sequence of response procedures through a parent cell which is split into binary offspring cells; this is termed the cell-sequence-cycle with a total time of 24 hours. It consists of five phases, as shown in Fig. 1(a), Gap1

(G1) 8-10 hours, DNA synthesis (S) 6-8 hours, Gap2 (G2) 4-6 hours, Mitosis (M) around 4 hours and Quiescence (G0) silent mode. Furthermore, the mammalian cell has substantial impact on living cell dynamics, involving cell proliferation with differentiation [1]. However, mammalian cells usually stay in the early state or resting state, either Quiescence the G0 phase or initial G1phase, but the cell cycle developments to S phase further than the check point when actual growing influences motivate a cell necessarily. After DNA duplication through the S phase, the cell cycle developments complete the G2 phase to the final phase called the M phase. At the end of the cell cycle, specifically through M phase, the cell is necessarily separated into two new cells, called daughter cells. It signifies the complete progression process in the cell cycle as illustrated in Figure 1(a), (b). Also, this progression process is controlled by several complex networks. These networks enclose several biochemical species such as genes and proteins [2].

A mammal cell is commonly damaged and harmed by different resources like ultraviolet (UV)-irradiation, also ionization-radiation (IR), or other toxic chemical elements that are able to influence and cause breaks inside double-stranded DNA. This leads to DNA damage, and simulates an exceptional signal in the cell. Precisely, this DNA damage signal fires a DNA damage signaling pathway. The signaling pathway cooperates with the cell cycle controlling system to tentatively stop the cell cycle evolution in order to repair damaged DNA. Naturally, DNA damage is organized through a sub-network with five components, as shown in Fig. 2, and they cooperate over these steps (1) double significant elements at the launch are activated such as Ataxia telangiectasia mutated (ATM) and Rad3-related (ATR) protein kinases are activated via DNA damage, (2) ATM and ATR prompt p53 and checkpoint kinase 1 (Chk1), (3) initiated p53 stimulates the synthesis of p21, (4) then p21 prompting cell cycle halt. Accordingly, the signaling pathway for DNA damage takes straight action on the cell cycle arrangement mechanism, supporting cellular homeostasis and genetic constancy. Besides, any cell that has significant DNA damage might prompt apoptosis and perform planned cell death [3], [4].

The influence of DNA damage in mammalian cells is one great cause of human illnesses, and as such has gained much interest in research since the mid-1990s. DNA damage and oxidative stress are identifying factors for the source,

\*Corresponding Authors.

development, aetiology and progression of numerous different types of human disorders and diseases, such as cancer. Therefore, an abundance of present-day investigation in the DNA damage domain is dedicated towards sympathetic mechanisms and natural allegations of harmful DNA. This harmful DNA can go through alterations, such as mutations on its genetics; these types of mutations ultimately prime the expansion of tumors. DNA damage is also concerned in the growth of further prevalent human illnesses ranging from neurodegenerative disorders, such as Alzheimer’s illness, to chronic obstructive pulmonary disease (COPD). Further, they have also been linked to diverse illnesses, such as pulmonary illnesses, brain injury and other chronic inflammation related to disorders [5].

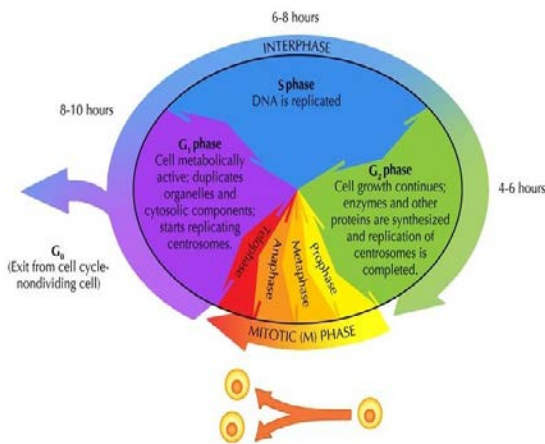
Essentially, biological discoveries have shown that mammalian cells are able to approximate the intensity-concentrations of DNA damage and choose a suitable cell fate, like applying DNA repair; otherwise, cell cycle arrest, or apoptosis death. Nevertheless, it is uncertain the manner in which a cell decides the suitable cell destiny. A confirmation of the affiliation among the intensity (proteins concentration) for DNA damage and the energetic behavior of the biochemical elements implicated in cell cycle controlling techniques and the signaling pathway for DNA damage is crucial for clarifying the techniques of cell destiny purposes.

DNA damage has a number of intensity-concentration levels; these levels of DNA damage in malignant cells and in further diseased cells are significant in the assessment of the lesion stages that appear in normal cells. A wealth of laboratory research has been concerned with distinguishing and comprehending DNA damage levels and DNA repair capacity, as well as the techniques employed through mutually abnormal and normal cells. In addition, since certain significant diseases, such as cancer, are the essential reason of premature mortality over the globe, there is a predominance and rapid assertiveness of research on illustrative DNA damage in cancer cells [6]. Mainly, as mentioned before, clarifying the DNA damage level will be helpful for treatments and research purposes, even if there is a scarcity in the available data. Consequently, this article aims to present a novel artificial deep learning model to classify and predict DNA damage levels based on available DNA damage intensity-concentration levels from a bench mark model. The bench mark model delivers a novel dataset for DNA signaling network and classifies the DNA damage into several levels. We rely on this dataset to train and test the novel proposed model to predict the weather of the DNA damage and classify them in several classes.

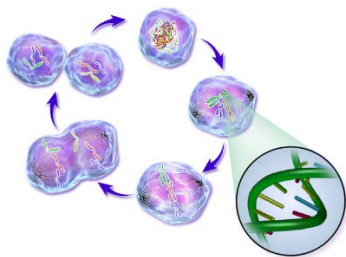
The research introduces and validates the novel model to predict and classify DNA damage levels. To achieve the goals, outcomes have been investigated with DNA damage datasets. The research objectives and contributions are represented as follows:

- 1) This research aims to propose a novel deep learning model by employing an Attention – Long Short Term Memory.
- 2) Experiments are to be applied on DNA damage datasets.
- 3) The ATT-LSTM deep learning classifier for DNA damage is to be employed, and the efficiency of the ATT-LSTM deep learning classifier is to be determined.
- 4) The developed framework not only solves the long-term dependence problem of prediction effectively, but also enhances the interpretability of the prediction methods established on the neural network.

The paper is structured as follows. Section II offers a literature review. Section III encloses the utilized method and model architecture and implementation. Section IV encompasses the results and investigation. Section V presents the conclusions.



(a) The Main Phases inside Mammalian Cell Cycle System.



(b) Illustration for Mammalian Cell Progression and Division Process

Fig. 1. Cell Cycle System Stages and Progression for Mammal Cell.

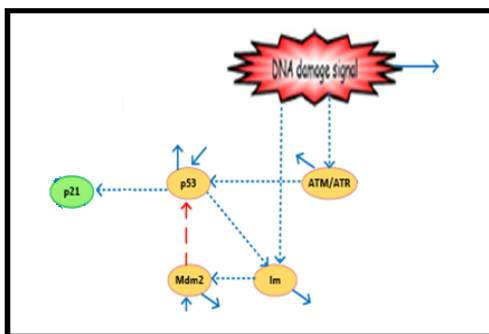


Fig. 2. The Main Elements for DNA Damage Signaling Pathway.

## II. RELATED WORK

Declaration and quantification of DNA damage is an actual substantial topic in biological and biomedical study areas, which requires further influential and active methods. Defining the DNA damage level is a significant point to decide the fate of the cell, such as if the cell recovers the DNA damage, or kills itself, or develops into an abnormal cell and forms into a serious diseases such as cancer [7]. Besides, the defining level can help to get more insight over drug treatment experiments. Several attempts have been made to classify DNA damage levels. For instance, numerous classical machine-learning methods have been employed in classifying the data that were related to gene expression, involving Fisher linear discriminant analysis [8], decision tree, k nearest neighbor [9], multi-layer perceptron [10], support vector machine [11] [12], boosting, and self-organizing map [13]. In addition, concerning clustering gene expression data, various machine learning techniques have been utilized [14]; they include hierarchical clustering [15], graph theoretic approaches [16] [17] and self-organizing map [18]. Concerning disease and treatment for DNA damage also, another attempt is available in the literature based on use of the machine learning classifiers on illness datasets, like Leukemia disease dataset, Lymphoma malignance data set and colon tumor dataset. Researchers have also attempted to explore many features by utilizing classical methods, like multi-layer perceptron neural network, k-nearest neighbor, structure adaptive \_SOM- self organizing map and SVM (support vector machine); these have been employed for classification [19]. In addition, they have joined the classifiers to increase the performance of classification. The experimental consequences indicate that the ensemble with some basic classifiers produces the greatest classification rate on the benchmark dataset.

Other researchers have established an SVM classifier exactly for mtDNA missense variants [20]. Therefore, in the process which is associated in the training and validation of the model, they employed 2,835 mtDNA damaging and neutral amino acid replacements. In the abovementioned dataset, each instance is well-defined through a fixture of three attributes created on evolutionary preservation in Eukaryote modified amino acids. Consequently, the proposed classifier achieved better than other web-available tested predictors. However, lately, a Deep learning model has been offered [21]. The model is based on a weak label learning method; they used this method to investigate the whole slide images (WSIs) of Hematoxylin besides Eosin (H&E). Their occupation was Self-supervised pre-training technique and heterogeneity aware deep Multiple Instance Learning (DeepSMILE) and they engaged it on cancer tissue images. Their model improvements recommended the genomic label classification performance without collecting larger datasets. There is also a deep learn pipeline based open source, called FociNet [22]. It is interested in image classification and was established to mechanically segment full-field fluorescent images and divide DNA damage of each cell. The outcome from the model indicated that FociNet reached satisfying performance in classification. Since it classifies a solitary cell in a normal, injured, or no signaling (no fusion-protein expression) state, and it also shows exceptional matching in the assessment of

DNA damage, contingent on fluorescent foci images from different imaging platforms [23]. Evaluation of the performance of convolutional Neural Network was done to examine the amount of DNA damage by means of comet assay images and was matched to further approaches in the literature. The novelty of their work was employing convolutional Neural Network as a novel scheme to classify the comet objects on segmented comet assay inside the images. Additionally, numerous deep learning models were applied on DNA damage images [24] [25] [26]. However, almost all the available deep learning models in the available literature are based on image datasets for DNA damage, while few available deep learning models are based on DNA damage intensity – concentration datasets; this is due to a scarcity in experimentally observed data concentration datasets. Therefore, while it is challenging to envision these complicated relationships using only DNA damage, investigators can systematically confirm these associations with a mathematical-numerical model that incorporates data from experiments toward a kinematic mathematical model which includes the cell cycle regulation techniques with the DNA damage signaling pathway. Various scientists have developed valuable kinetic mathematical models. These models are associated with the cell cycle regulation mechanism to estimate the exchanges of natural species [1], [7], [27], [28], [29], [30], [31].

A mathematical model was proposed as a benchmark model of the DNA damage-signaling pathway and mimic cell fate selection [30]. The outcome from the novel model was that it offers a dataset for the DNA damage signaling pathway. This dataset exposes the proteins' concentration levels and activities to deal with DNA damage level. For instance, the researchers presented the DNA damage signaling pathway-proteins set-concentrations without DNA damage [30]. These observations from the delivered dataset qualitatively match with biologically appropriate facts. In addition, diverse intensities of DNA damage were found, such as Low-damage, Medium-damage, High-damage, and Excess-damage. These aforementioned DNA damage levels bear a resemblance to actual DNA damage, and are triggered by several values such as 100, 200, 400, and 800 J/m<sup>2</sup> doses of UV-irradiation. Further explanation will be clarified in the dataset preparation and analysis section and how we utilized this dataset to train and test the proposed artificial deep learning model to predict the DNA damage level into several classes.

## III. PROPOSED WORK AND OVERVIEW OF SYSTEM ARCHITECTURE (ATTENTION – BASED LSTM) – AT-LSTM MODEL

Currently, there is no effective computational model, neither a machine learning nor a deep learning model that can be utilized to validate the influence of the intensity-concentration of DNA damage on cell cycle system progression. Consequently, the crucial contributions of this paper essentially comprise the following: we employ the attention model to effectively extract the features of DNA damage big and complex dataset and the LSTM layer in the proposed model performs additive interactions, which can help improve gradient flow over long sequences in training [32]. Matched with classical models, AT-LSTM can

competently maintain and work with non-stationary sequences and detect the nonlinear relationships [33]. Furthermore, compared with deep learning models like RNN, the AT-LSTM can avert the long-term dependence issues and give rise to superior interpretability [34]. The mechanism for the attention in the proposed model makes it simple to recognize how the information in the input sequence influences the final created sequence through the model output process [35]. This might assist in discovering the interior operation mechanism of the model and debug certain precise inputs and outputs. Further, the experimental results on DNA damage datasets determine that AT-LSTM accomplishes more enhanced tasks than standard models.

1) *The architecture of the AT-LSTM model:* The proposed attention-based LSTM (AT-LSTM) model for DNA damage dataset multiclass prediction comprises two parts: the attention model and the LSTM deep learning model. The attention mechanism is able to adaptively choose the furthest related input features and provide higher weights to the corresponding original feature sequence. Then and there, we utilize the outcomes of the LSTM deep learning model as input for the attention model to predict the DNA damage level and assign it to several classes.

2) *LSTM model:* For a stated input raw,  $X = (x^1, x^2, \dots, x^n)^T = (x_1, x_2, \dots, x_m) \in R^{(n \times m)}$ ,  $n$  represents the numeral of feature orders -sequences,  $m$  stands for the length of the window.  $x^k = (x_1^k, x_2^k, \dots, x_m^k)^T \in R^m$  is utilized to denote a sequence (vector) of length  $m$ . For biological DNA damage, this sequence can be a protein concentration measurement for the sub-network, which represent the DNA signaling network. We use  $x_t = (x_t^1, x_t^2, \dots, x_t^n)^T \in R^n$  to represent a set-group of vectors of  $n$  features at time  $t$ . Long Short-Term Memory (LSTM) model is declared as follows: Let  $x_t, h_t$  and  $C_t$  stand for the input, control state, and the cell state on time step  $t$ . Delivering a sequence of inputs  $(x_1, x_2, \dots, x_m)$  the LSTM calculates the group of sequence  $(h_1, h_2, \dots, h_m)$  and the C-sequence  $(C_1, C_2, \dots, C_m)$  as follows:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad (1)$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad (2)$$

$$c_t = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \quad (3)$$

$$C_t = f_t * C_{t-1} + i_t * c_t \quad (4)$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad (5)$$

$$h_t = o_t * \tanh(C_t) \quad (6)$$

such that each equation has a set of special symbols, and identify several functions. For occurrence,  $\sigma$  represents the function of logistic sigmoid,  $*$  is a component wise multiplication, and  $C_t$  is the weather of the cell that is required to be changed. Also,  $W_f, W_i, W_c, W_o$  and  $b_f, b_i, b_c, b_o$  are a set of parameters for the model. Besides, these parameters can be learned over the processing. Additionally,  $f_t, i_t$  and  $o_t$  are likewise christened as a gate for the forgotten, along with an input gate and output gate. In actual fact, the architecture for

the LSTM unit includes a memory cell, this mean that every LSTM unit that contains a memory cell has state  $C_t$  at time  $t$ , which is structured by the three overhead gates.

3) *The attention model:* A significant part of human artificial is that it does not directly contract with all feedbacks from the outside world. As a substitute, human's first attention is on the significant sections to acquire the information they require. Correspondingly, the significance of several proteins concentrations in the biological data set is also different, big, and complex, and the other may be critical. It is also essential to emphasize key features first and remove repeated features. Accordingly, with the operative information inspired through the overhead information, we propose an attention model, and this model can apply the optimization part for the input feature sequence in DNA damage level prediction. An attention mechanism [35] can be defined as mapping an enquiry. Moreover, a set of key-value couples to an output, and similarly, the components in the system such as keys, query, values, including output are all defined as vectors. The outcome of the model is calculated as a weighted sum for the values, where the weight given to every value is calculated through a function related to the compatibility for the query with the equivalent key, as shown in Fig. 3.

The method of producing attention weights and the new input features established on attention is illustrated in Fig. 3. In the first fragment,  $x_t$  maps to  $h_t$  through the following.

$$h_t = f_1(h_{t-1}, x_t) \quad (7)$$

where the non-linear activation function is represented by  $f_1$ , while  $h_t \in R^s$  stands for the hidden state on time  $t$ , and  $s$  indicates the size of the hidden state. LSTM is implemented as  $f_1$ . The main aim for this implementation is to evade the long-term dependence problem, which typically arises in data prediction.

In the second fragment, we generate an attention mechanism by using specifically the deterministic feature in the attention model. For an exact feature sequence like  $x^k = (x_1^k, x_2^k, \dots, x_m^k)^T \in R^m$ , by relying to the aforementioned hidden state  $h_{t-1}$  and the cell state  $C_{t-1}$  in the LSTM unit, we express

$$\alpha_t^k = v^T \tanh(W_1 \cdot [h_{t-1}, C_{t-1}] + W_2 x^k) \quad (8)$$

$$\beta_t^k = \text{softmax}(\alpha_t^k) = \frac{\exp(\alpha_t^k)}{\sum_{i=1}^n \alpha_t^i} \quad (9)$$

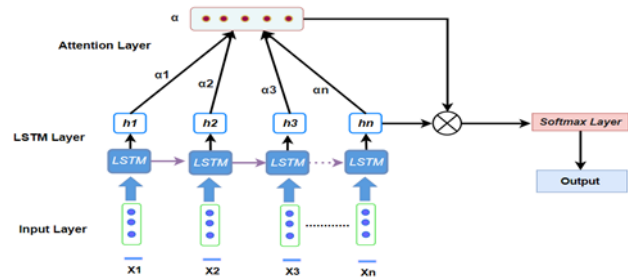


Fig. 3. The Architecture of the Proposed Attention-LSTM Model.



The vector  $v$  and the two matrices  $W_1, W_2$  signify the learning abled parameters of the proposed model. The vector  $a^k$  has a length called  $m$  and its  $i$ -th item measures the significance of the  $k$ -th input feature sequence at time  $t$ . The aforementioned items must be normalized through softmax.  $\beta^k$  represents the weight in attention, which encloses a score, and the score shows the amount of attention that should be put on the  $k$ -th feature sequences. We are able to likewise acquire the outcome of the attention model at time  $t$ , i.e., the sequence of the weighted input feature named as  $z_t$  can be presented as follows:

$$z_t = (\beta_t^1 x_t^1, \beta_t^2 x_t^2, \dots, \beta_t^n x_t^n)^T \quad (10)$$

$x_t$ , in the equations from (1) to (7) swapped via a new calculated  $z_t$  to keep up the attention model. However, classical prediction frameworks that enclose recurrent neural networks usually utilized dataset input features as input, besides treating all input feature sequences in an equivalent fashion. Nevertheless, the recently acquired  $z_t$  can pay further attention to the particular input feature sequence, mining the key feature sequences efficiently, and based on attention weight, we reduced the influence of the redundant feature sequences. Hypothetically, there would be an improvement in prediction exactness with  $z_t$  as the input to the softmax layer.

#### IV. RESULTS AND DISCUSSIONS

##### A. Data Analysis and Simulation

This section explores how we apply experiential research on data sets with an aim to elucidate the validity of our DNA damage level prediction framework. First, we will introduce the dataset that was utilized in training and testing of the proposed model. We relied on available biological datasets related to DNA damage signaling pathways, which plays a crucial role in DNA damage in mammalian cell systems [30]. This Biological dataset consists of 15000 records intensity – concentration-level for a five-proteins set which control DNA damage. The DNA damage signaling pathway- proteins set-concentrations without DNA damage was previously presented [30]. These observations from the delivered dataset were qualitatively analyzed with biologically appropriate facts. In addition, diverse intensities of DNA damage were performed, such as Low-damage level, Medium-damage level, High-damage level, and Excess damage level. Herein, we studied and analyzed the aforementioned dataset and proposed a novel model which consists of an attention based on long short term memory- Neural Network named as LSTM (AT-LSTM) model for DNA damage multi-classification prediction.

##### B. Dataset Analysis

Researchers have assembled a new kinetic based mathematical – ordinary differential equations (ODE's)-model that assimilates the G1/S in cell cycle system models, and they measured compatibility to the biological credibility of the suggested model by confirming numerous mathematical mimicry time progression courses of the intensities of individual biochemical elements [30]. Furthermore, they as well quantitatively recognized the intensity – concentration level of DNA damage and provide experimentally observed data.

Certainly, when DNA damage has occurred, numerous protein kinases are involved at the location of damage and launch a special signaling pathway that forces cell-cycle to be arrested. The chief kinase at the damage location is ATM/ATR, which is activated and established on the type of damage and another protein of the gene regulatory protein p53 is also triggered. Mdm2 usually connects to p53 and stimulates its ubiquitylation and destroys the proteasomes. Phosphorylation of p53 stops its binding to Mdm2; consequently, p53 becomes accumulated to maximum levels and inspires transcription of the gene that encrypts the protein p21 and arresting of the cell in G1 [30].

Mainly, in the evolution process for the model, first, we extracted the observation from a base model deprived of DNA damage (DDS = 0) to get the time course for selected cell cycle regulators. Second, we extracted from an expermental dataset of the benchmark model the required obseravtion with four diverse levels of DNA damage: (Low-damage) with DDS = 0.002, (Medium-damage) with DDS = 0.004, (High-damage) with DDS = 0.008, and (Excess-damage) with DDS = 0.016. If DNA damage has certainly not arisen, the p21 with p53 stop over instead, and with a low level, as illustrated in Fig. 5 and 6. DNA damage drives p53 activation which prompts p21 [2].The character of p21 is to prevent the activity through inhibition of phosphorylation of Rb [1].

With the elimination of DNA-damage, [2]p53 and Mdm2 have a negative feedback loop which is completely reinstated, and p53 returns back to a low-slung level. The reduction in p53 decreases the scale of p21, as shown in Fig. 4 and 5 when DDS=0.004. Likewise, Figure 4 shows all the protein tensity for the protien P21 response in all DNA damage states. For instance, it shows the values for the P21 tensity in the case without DNA damage occurrence as a light-blue line. It provides roughly 3000 instances. On the other hand, it shows the concentration values for the P21 when DNA damage occurs,; for instance, in case of Low DNA damage, the P21 can be presented in an orange line, medium damage with a gray line, and a high DNA damage level P21 values and behaves to recover the DNA damage represented in a yellow line, while the P21 response in extreme DNA damage is represented in a dark blue line. We have to be aware about the values for each figure, specifically that each element in each DNA damage state has roughly 3000 different instances in response to DNA damage recovery.

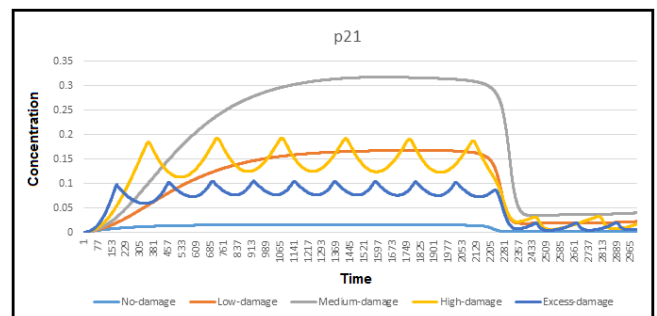


Fig. 4. Time Courses of P21 responses with and without DNA Damage over the Simulation of Mammalian Cell.

With respect to High DNA damage with a rate of  $DDS=0.008$ , the time progressions of p21 and p53 are revealed in Fig. 5 and 6. The p53 is activated and presented in oscillation behavior, which were in settlement by means of those previously experimentally detected [30, 6, 29]. Furthermore, when DNA damage is presented, the DNA-damage signal in the sequence triggers p53 instead of Mdm2. The triggered p53 similarly can stimulate the synthesis of p21 which acts as an inhibitor. Meanwhile, p21 stops the phosphorylation of Rb. Fig. 6 indicates all the protein tensity for the protein P53 response in all DNA damage states. For example, it explores the values for the P53 tensity in the case without DNA damage incidence, the light blue line. On the other hand, it displays the concentration values for the P53 when DNA occurred, such as in the Low DNA damage, in which case the P53 is presented with an orange line, and the medium damage is represented with a gray line. In the case of high DNA damage levels, P53 values and behaves to recover the DNA damage that is represented in a yellow line, while the P53 responds in extreme DNA damage cases, represented in a dark blue line.

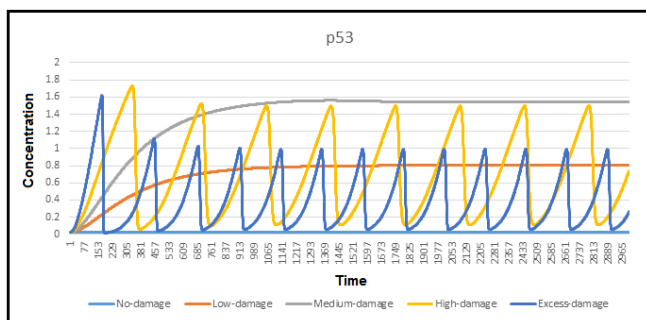


Fig. 5. Time Courses of P53 with and without DNA Damage over the Simulation of Mammalian Cell.

Fig. 6- 8 illustrate the responses of Mdm2, ATM/ATR and Im respectively in cooperating to handle the DNA damage cases. Each figure, as explained before, shows how the values in protiens tensity of each element will change during the DNA damage, whether it occurred or not. As demonstraed before, we have to be alert that the values for each figure of each ellement in each DNA damage state has around 3000 different instances in response to DNA damage recovery.

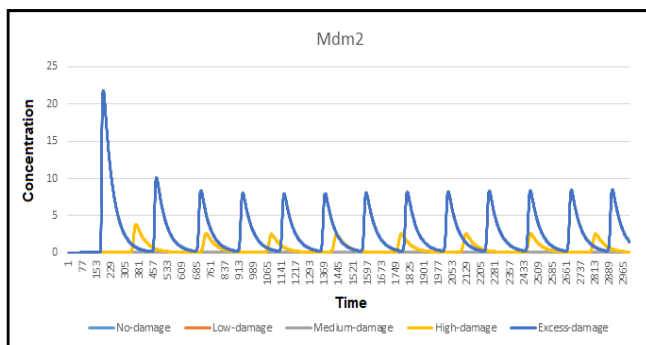


Fig. 6. Time Courses of Mdm2 with and without DNA Damage over the Simulation of Mammalian Cell.

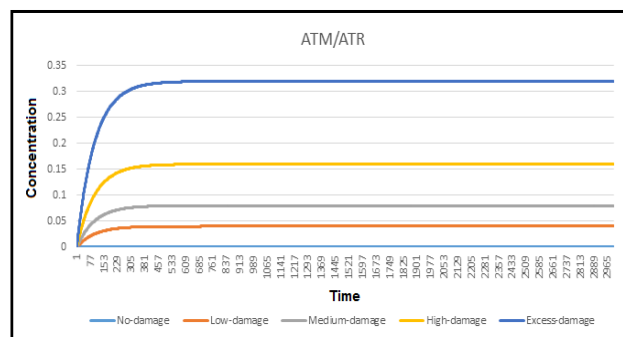


Fig. 7. Time Courses of ATM/ATR with and without DNA Damage over the Simulation of Mammalian Cell.

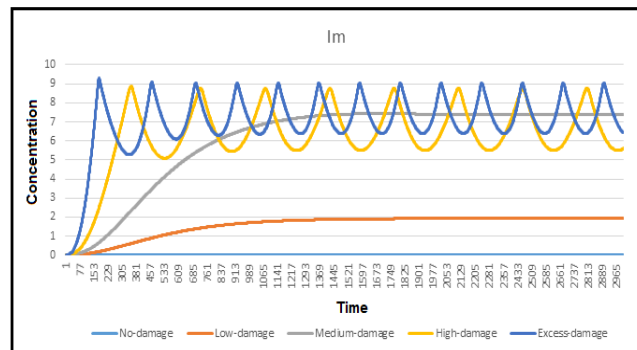


Fig. 8. Time Courses of Im with and without DNA Damage over the Simulation of Mammalian Cell.

Herein, we revealed a brief examination of biological background, specifically cell cycle, with more deliberate focus on DNA damage pathways as a complicated system. As demonstrated before, the novel proposed deep learning Attention based LSTM model is trained and tested depended on the obtained dataset delivered by [30].

### C. Experiments

We performed several experiments on the proposed attention-based LSTM model for DNA damage classification. The proposed model was trained on 12000 samples and was tested on 3000 samples. The dataset is in-of-domain for DNA damage classification and the tested dataset that was used is also from the same dataset. There are three cross-validation methods which are often employed to evaluate the success rate of the predictor; namely, the K-fold cross validation, sub-sampling and jackknife test. The Jackknife test is the least arbitrary and most objective, and it has been mostly assumed by researchers to inspect the quality of diverse predictors. This method is source and time consuming. Therefore, in this paper we utilized an early stopping choice to elude a model's overfitting through setting the patience option to three epochs, and we utilized k-fold cross validation where K was set to 1, such that a single train/test split is generated to evaluate the attention-based LSTM model for DNA damage classification.

### D. Training

The framework developed by Keras and Python was used to train the attention-based LSTM model for DNA damage classification. For the classification task, SGD optimization algorithm was used with learning rate values set to 0.01 and



momentum set to 0.0, and the model's batch size set to 6. The model initially incorporated 331,525 parameters. The model used 256 LSTM units and the attention dimension was set to 255. Also, the proposed model size proved to require 1.7 seconds per epoch for the classification task. The training data was randomly shuffled at each epoch for the classification task. The proposed attention-based LSTM model for DNA damage classification task was trained to minimize the categorical-crossentropy validation loss for the DNA damage classification task while maximizing validation accuracy for the same classification task.

### E. Results Investigation

On performing experiments, it can be assured that the researchers have done extensive experiments on the attention-based LSTM model by testing different hyper-parameters. The proposed model was also experimented using three different configurations, such as BiLSTM with Attention layer, LSTM with Attention Layer and LSTM with Attention and Dropout layers. The classification performances measured will be listed in an accuracy score automatic metric evaluation. The results in Table I show the efficiency of the proposed Attention-Based LSTM model for DNA damage classification. It can be noticed from Table I that the proposed model obtained an excellent result when we used LSTM with the attentional approach, such that the model obtained an accuracy of 93.43. In comparison to other configurations listed in Table I, the proposed model (LSTM with Attention) obtained a better accuracy than the other configurations. These results suggest that the proposed model is effective and accurate in classifying DNA damage in a validation dataset. Also, as seen from Table I, the BiLSTM with attention configurations has obtained a competitive result, such that it obtained an accuracy of 93.13, which indicates that the BiLSTM performs very well in classification tasks. Adding a dropout layer to the model's design negatively impacts the model's performance and quality, such that the model obtained an accuracy of 75.43, as illustrated in Table I. Therefore, it can be summarized that the proposed model outperforms the models that used BiLSTM and Dropout layer. More importantly, results presented in Table I and Fig. 9 show the performance of the model that exploited the attention approach, and LSTM is higher than the other models. In addition, as shown in Fig. 10, it can be seen that the error on the training data decreases as the learning continues, and at the same time, the error of actual validation data decreases as the training continues, and this pattern proves that the proposed AT-LSTM model is not facing the problem of overfitting.

Moreover, as shown in Fig. 11, plot of accuracy, we found that the model is trained very well, as the trend for accuracy on both training and test datasets were still rising from epoch 100 till epoch 213, and this is an indication of the proposed model's performance and accurate classification.

Fig. 12. (a,b,c,d,e) illustrates how the AT-LSTM model classifies the responses of P21, P53, Mdm2, ATM/ATR and Irf, respectively in cooperating to handle the DNA damage. Each figure shows how the values in protein density of each element changes during DNA damage, and how it occurred in each class. The x-axis represents the diverse classes for the

DNA damage level while the y-axis represents the amount of each protein concentration during the DNA damage.

TABLE I. EXPERIMENTAL RESULTS

Model Configuration	Accuracy	Number of epochs
BiLSTM +Attention	93.13	228
The proposed model (LSTM with Attention )	93.43	213
LSTM + Attention + Dropout	75.43	71

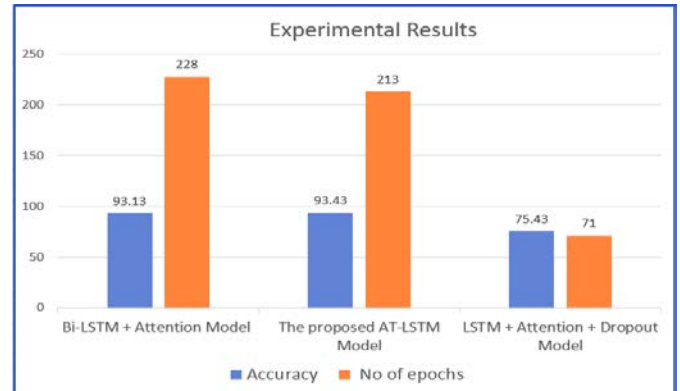


Fig. 9. Model Accuracy within Number of Epochs.

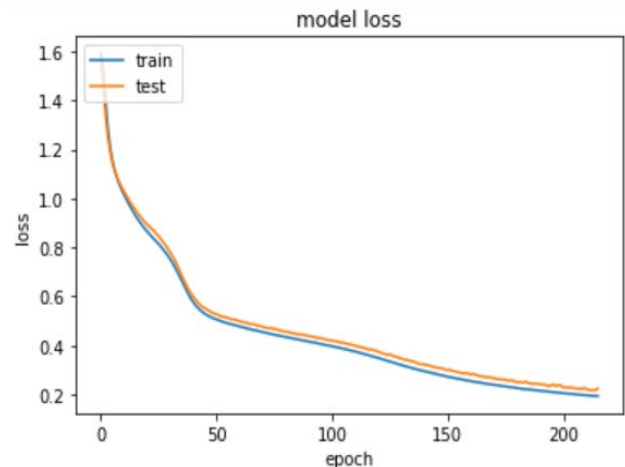


Fig. 10. Model Loss.

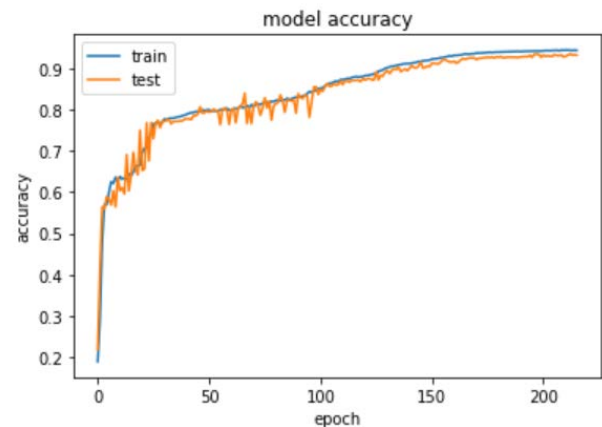
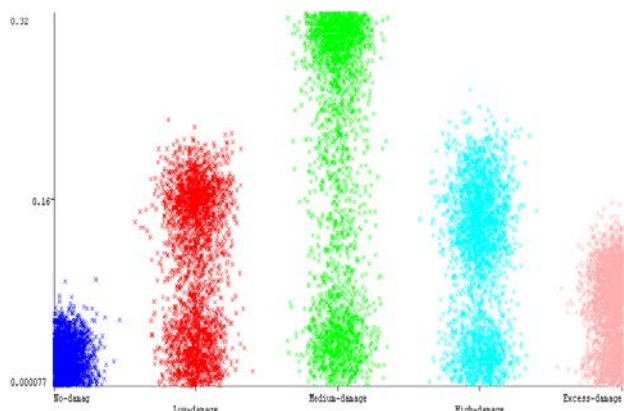
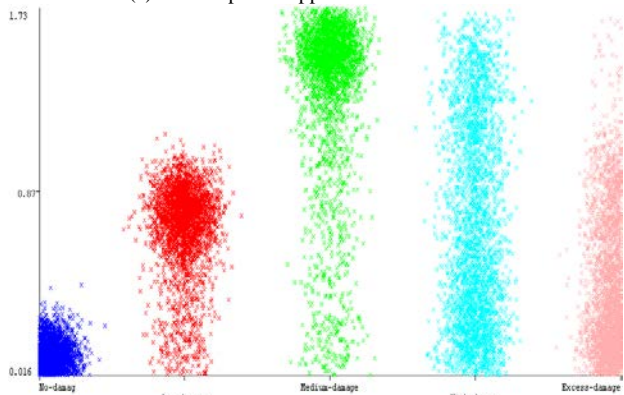


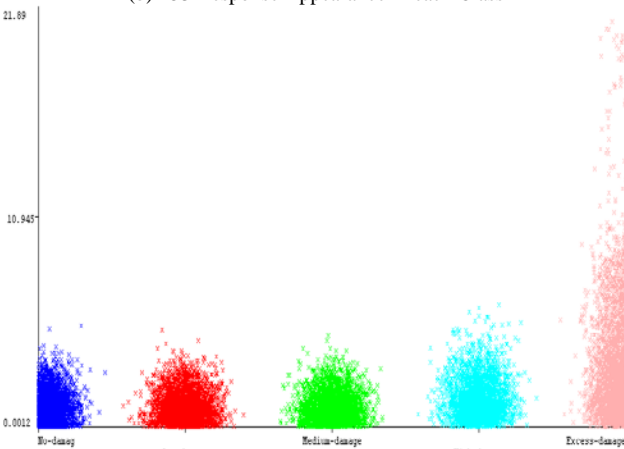
Fig. 11. Model Accuracy.



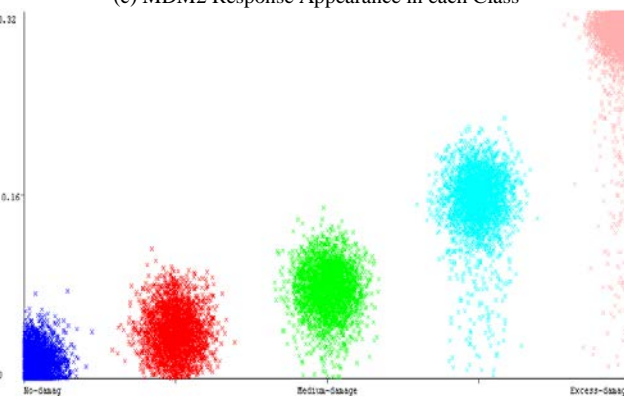
(a) P21 Response Appearance in each Class



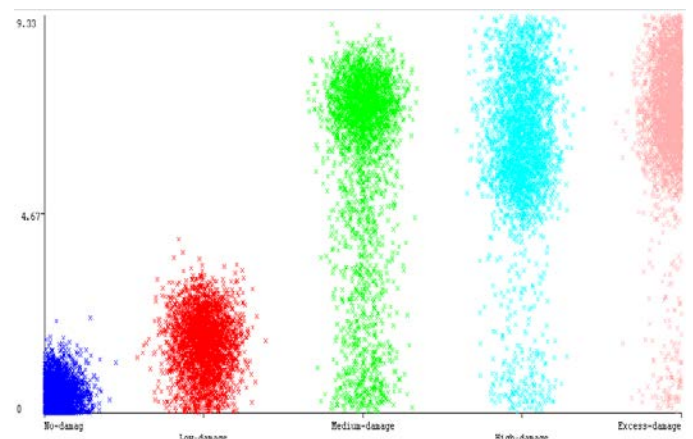
(b) P53 Response Appearance in each Class



(c) MDM2 Response Appearance in each Class



(d) ATM/ATR Response Appearance in each Class



(e) Im Response Appearance in each Class

Fig. 12. AT-LSTM Model Classification for each Protein in Response to DNA Damage Level.

## V. CONCLUSION

DNA damage in mammalian cells causes genetic illnesses and a diversity of cancers. Therefore, more investigation and analysis can help the therapeutic process. Almost all classification prediction models that are used to explore DNA damage are based on DNA damage images, while few studies and models are based on DNA damage intensity. In this paper, we developed a novel deep learning model; in essence an Attention-based LSTM model to perform classification tasks of DNA Damage Levels. The proposed model was able to overcome other models and obtained an accuracy of 93.43%. These results confirm that the proposed model is effective and accurate in predicting and classifying DNA damage on a validation dataset. The attention approach was able to extract the complex features from the dataset and enhanced the proposed model quality and performance. The proposed model is considered as a novel work since AT-LSTM has never been applied in the DNA damage field, and can be employed to assist the investigation and studies of DNA damage since it provided a promising prediction of results.

## REFERENCES

- [1] Y. H. H. O. M. H. T. Tashima, "Prediction of key factor controlling G1/S phase in the mammalian cell cycle using system analysis," *Biosci.*, pp. 106 (4), 368–374., 2008.
- [2] K. I. Kohn, "Molecular interaction map of the mammalian cell cycle control and DNA repair systems." *Mol. Biol. Cell*, pp. 10, 2703–2734, 1999.
- [3] G. H. V. Li, "p53-dependent DNA repair and apoptosis respond differently to high- and low-dose ultraviolet radiation," *Br. J. Dermatol*, pp. 139, 3–10., 1998.
- [4] W. K. B. Roos, "DNA damage-induced cell death by apoptosis. Trends," *Mol. Med.*, pp. 12 (9), 440–450., 2006.
- [5] N. P. M. Sharbel Weidner Maluf, "DNA Damage and Oxidative Stress in Human Disease," *BioMed Research International*, 2013.
- [6] M. D. Bryant C Nelson, "Implications of DNA damage and DNA repair on human diseases," *Mutagenesis*, , pp. Volume 35, Issue 1, 2020.
- [7] H. T. Y. K. Y. I. K. H. T. E. Y. O. Hamada, "Sophisticated framework between cell cycle arrest and apoptosis induction based on p53 dynamics." *PLoS ONE*, pp. 8, 28–35, 2009.
- [8] S. F. J. a. S. T. P. Dudoit, "Comparison of discrimination methods for the classification of tumors using gene expression data," Department of Statistics, University of California, Berkeley, p. Technical Report 576, 2000.

- [9] L. W. C. R. D. T. A. a. P. L. G. Li, "Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the GA/KNN method." *Bioinformatics*, pp. 17(12):1131-1142, 2001.
- [10] J. W. J. S. R. M. S. L. H. L. M. Khan, "Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks," *Nature Medicine*, pp. 7(6):673-679., 2001.
- [11] T. S. C. N. D. N. B. D. W. Furey, "Support vector samples using microarray expression data," *Bioinformatics*, pp. 16(10):906-914. , 2000.
- [12] M. P. S. G. W. N. L. D. C. N. Brown, "Knowledge-based analysis of microarray gene expression data by using support vector machines," *Proc. of the Natl. Acad. of Sci. USA*, pp. 97:262-267, 2000.
- [13] T. R. S. D. K. T. P. H. C. Golub, "Class discovery and class prediction by gene-expression monitoring." *Science*, pp. 286:531-537, 1999.
- [14] R. a. S. R. Shamir, "Algorithmic approaches to clustering gene expression data," *Current Topics in Computational Biology*, MITpress, 2001.
- [15] M. B. S. P. T. B. P. O. a. B. Eisen, "Cluster analysis and display of genome-wide expression patterns," *Proc. of the NatlAcad. of Sci. USA*, pp. 95:14863-14868, 1998.
- [16] E. S. A. L. J. M.-E. S. Hartuv, "An algorithm for clustering DNA fingerprint," *Genomics*, pp. 66(3):249-256, 2000.
- [17] A. B. L. F. N. N. Ben-Dor, "Tissue classification with gene expression profiles," *Journal of Computational Biology* , pp. 7:559-584, 2000.
- [18] P. H. C. Tamayo, "Molecular classification of cancer: Class discovery and class prediction by gene-expression monitoring," *Science*, pp. 286:531-537, 1999.
- [19] S.-B. C. a. H.-H. Won, "Machine Learning in DNA Microarray Analysis for Cancer Classification," *DBLP*, 2003.
- [20] A. G.-S. Antonio Martín-Navarro, "Machine learning classifier for identification of damaging missense mutations exclusive to human mitochondrial DNA-encoded polypeptides," *BMC Bioinformatics*, p. Article number: 158 (2017), 2017.
- [21] E. G. I. N. H. M. H. J. T. Yoni Schirris, "DeepSMILE: Self-supervised heterogeneity-aware multiple instance learning for DNA damage response defect classification directly from H&E whole-slide images," *Electrical Engineering and Systems Science*, 2021.
- [22] D. X. 1. R. Z. 2. L. Z. Xuechun Chen 1, "Deep-Learning-Assisted Assessment of DNA Damage Based on Foci Images and Its Application in High-Content Screening of Lead Compounds," *publmid*, pp. :14267-14277., 2020.
- [23] Y. B., E. Ş. Umit Atila, "Classification Of Dna Damages On Segmented Comet Assay Images Using Convolutional Neural Network," *Computer Methods and Programs in Biomedicine*, p. 186:105192, 2019.
- [24] R. H. M. S. a. D. W. G. Christoph Sommer, "A deep learning and novelty detection framework for rapid phenotyping in high-content screening," *Molecular Biology of the Cell*, pp. Vol. 28, No. 23, 2017.
- [25] M. K. Y. S. F. K. O. Jianzhu Ma, "Using deep learning to model the hierarchical structure and function of a cell," *Nature Methods*, pp. 15, pages290–298 , 2018.
- [26] D. M. a. S. Vadivazhagu, "Predicting DNA Damage in Fluorescent Imaging," *Worcester Polytechnic Institute*, 2020.
- [27] Z. W. J. M. W. Qu, "Regulation of the mammalian cell cycle:a model of the G1-to-S transition," *Cell Physiol*, pp. 284, 349–364, 2003.
- [28] B. T. J. Novak, "A model for restriction point control of the mammalian cell cycle," *Biol*, pp. 230, 563–579, 2004.
- [29] A. B. D. C. K. N. B. T. J. Csikasz-Nagy, "Analysis of generic model of eukaryotic cell-cycle regulation.," *Biophys*, pp. 90 (12), 4361, 2006.
- [30] H. H. Y. E. M. O. Kazunari Iwamoto, "Mathematical modeling of cell cycle regulation in response to DNA damage:," *BioSystems*, p. 384–391, 2011.
- [31] H. K. D. S. S. Ling, "Robustness of G1/S checkpoint pathways in cell cycle regulation based on probability of DNA-damaged cells passing," *Biosystems*, pp. 101, 213–221, 2010.
- [32] Z. Fu J, "Look closer to see better: recurrent attention convolutional neural network for fine-grained image recognition," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4476-4484, 2017.
- [33] J. L. Elman, "Distributed representations, simple recurrent networks, and grammatical structure," *Machine learning*, p. 195–225, 1991.
- [34] S. N. P. N. Vaswani A, "Attention is all you need," *arXiv preprint arXiv*., pp. 1706.03762., 2017.
- [35] C. D. L. H. a. A. M. Yoon Kim, "Structured attention networks," *International Conference on Learning Representations*, 2017.

# Estimation of Recovery Percentage in Gravimetric Concentration Processes using an Artificial Neural Network Model

Manuel Alejandro Ospina-Alarcón<sup>1</sup>, Ismael E. Rivera-M<sup>2</sup>, Gabriel Elías Chanchí-Golondrino<sup>3</sup>

Faculty of Engineering, Systems Engineering Program, Universidad de Cartagena, Cartagena de Indias, Colombia<sup>1,3</sup>

Faculty of Engineering, Pascual Bravo I.U, Medellín, Colombia<sup>2</sup>

**Abstract**—The concentrate process is the most sensitive in mineral processing plants (MPP), and the optimization of the process based on intelligent computational models (machine learning for recovery percentage modelling) can offer significant savings for the plant. Recent theoretical developments have revealed that many of the parameters commonly assumed as constants in gravity concentration modelling have a dynamic nature; however, there still lacks a universal way to model these factors accurately. This paper aims to understand the model effect of operational parameters of a jig (gravimetric concentrator) on the recovery percentage of the interest mineral (gold) through empirical modeling. The recovery percentage of mineral particles in a vibrated bed of big particles is studied by experimental data. The data used for the modelling were from experimental test in a pilot-scale jig supplemented by a two-month field sampling campaign for collecting 151 tests varying the most significant parameters (amplitude and frequency of pulsation, water flow, height of the artificial porous bed, and particle size). It is found the recovery percentage (%R) decreases with increasing pulsation amplitude (A) and frequency (F) when the size ratio of small to large particles (d/D) is smaller than 0.148. An empirical model was developed through machine learning techniques, specifically an artificial neural network (ANN) model was built and trained to predict the jig recovery percentage as a function of operation parameters and is then used to validate the recovery as a function of vibration conditions. The performance of the ANN model was compared with a new 65 experimental data of the recovery percentage. Results showed that the model ( $R^2 = 0.9172$  and  $RMSE = 0.105$ ) was accurate and therefore could be efficiently applied to predict the recovery percentage in a jig device.

**Keywords**—Empirical modeling; dynamic gravimetric concentration model; gravimetric concentration; machine learning for recovery percentage modelling; mineral processing

## I. INTRODUCTION

Recently more and more attention is paid to the methods of increasing the amount (yield) of gold concentrates obtained in separation by gravimetric processes that have an undesired gangue minerals (commercially worthless minerals) content. The purpose of gravity separation processes in jigs is to produce maximum amount of concentrate having desired gold content. This problem has been discussed during fifty last years in many research papers [1]–[32].

The mineral concentrate zone is a highly nonlinear process that requiring control; its parameters vary with time and

depend on the mineral feed rate and its size and density composition [30]–[34].

Although the use of neural network models is well established in the literature, it should be noted that such models, that depend on a large amount of data in the mineral processing industry is not very frequent, seeing the need to require such intelligent systems for further planning of design and optimization tasks, with which it is possible to understand, explain and test without the need to intervene in the real process [24], [35]. Machine learning models in general and artificial neural network models in particular, can be used to design and optimize control systems of concentrate discharge in jigs without the need to require a complex model that describes in detail the phenomena involved inside the equipment.

Regretfully, the lack of knowledge about all the phenomena that occur in this type of process is a frequent condition in practice in the mineral processing industry. Such situation occurs due to the low availability of phenomenological studies and due to some difficulties in modeling, inherited from past experiences; insufficient computational power led to the false appreciation that neural network models are complex. In addition, future research is needed to implement advanced control functions through the use of computer vision and multivariate data analysis. Such situations are directly reflected in the low availability of accurate models, which causes, for example, design problems in mining-metallurgical processes that must vary their operating conditions due to the heterogeneity of the ore to be processed. Today, it is possible to overcome such difficulties and use machine learning models such as neural networks to predict and describe the behavior of these processes, exploiting the capabilities of the model to analyze large amounts of data on the operational variables of the process while works with the process itself [32]–[34], [36].

Over the years, DEM and CFD models with a more detailed description of the gravimetric concentration phenomenon are being used more often [16], [29]. However, in spite of that progress, some challenges remain. For instance, one of the main goals that has not been accomplished so far is an agreed mathematical description of the percentage of mineral recovery (%R) and the operational parameters of the equipment (amplitude and frequency of pulsation, water flow, height of the artificial porous bed and size distribution). The %R turns out to be of paramount importance since it is directly

related to the amount of grams of gold per tons of ore concentrated (gpt): ideally, the gold percentage recovered from the concentrate stream should be equal to 100%.

Currently, the literature provides different ways to determine the %R, depending on the final application. As mentioned before, in some DEM and CFD models the authors use constant or linear approximations for %R [2], [13], [37]. Some other authors use empirical correlations describing the %R in terms of parameters mainly related to the geometry of the equipment and granulometric distribution of the mineral [31], [32], [38]. In spite of the many available ways to determine the %R in gravimetric equipment, the main observed concern about the above approaches is that they are only useful in the systems from which they were developed [2].

The fundamental purpose of this work is to apply and promote the use of artificial neural network models (ANN), in the analysis of gravimetric concentration processes for the design and control of equipment. The motivation for a work like this arises from the evident need to include the dynamic behavior of this type of process as fundamental elements in the design tasks in mineral processing engineering. In this regard, there is a large number of computational tools that already offer assistance for CAD (Computer Aided Design) without having written support that justifies such uses. Therefore, the inclusion of the concept of "machine learning" as the foundation of intelligent modeling is imperative, leading to the best use of the model as an analysis tool and support for process design and control tasks [2], [16], [24]. Since the existing works on gravimetric concentration equipment such as the jig, so far focus directly on a statistical analysis and describe the recovery percentages through experimental equations, and the mathematical models that currently describe the phenomenon are highly computationally demanding, making them unattractive for rapid tasks of optimization and control of equipment.

In this work, an accepted methodology for obtaining ANN is used in the modeling of the gravimetric concentration stage in the mineral processing industry. The efficient design and optimization tasks of the jig (which is the main equipment of the concentration stage in a mineral processing plant for gold extraction) require the process model to be able to study its behavior. Therefore, a good effort is dedicated here to the deduction and validation of an ANN model for the jig, counting on data available from a real pilot jig. This paper aims to develop a model to study the %R in a gravimetric pilot plant and its interaction with other internal process variables. First, an artificial neural network model is obtained to describe the dynamic behavior of the pilot plant. Then, the model is tested and assessed in terms of their ability to predict %R in the studied pilot plant.

The rest of the work is organized as follows: in Section II, a review of modeling in mineral processing is presented. In Section III, the ANN model is defined, mentioning the procedure for obtaining it, while in Section IV, said procedure is applied to the jig. Section V shows the simulation results of the obtained model and discusses its qualitative and quantitative validation, ending with a Section VI of conclusions and future work.

## II. RELATED WORKS

Jig is a gravimetric concentration equipment where mineral particles move in a flow of water pulsating, resulting at the end of the process in a stratification of particles of different densities and sizes. The stratification of the particles inside the jig occurs in a complex multiphase flow field. The particles are subjected to different hydrodynamic forces caused by the movement of the fluid, giving rise to different trajectories that depend on the velocity field of the fluid and particle properties. Various variables operations affect the motion of particles among which can be included the flow of fluid from water and mineral, the amplitude and frequency of pulsation of the fluid, among others [2].

Over the past decades, considerable efforts have been made to design and modify the jiggling process to increase gold recovery percentages and optimize processes [1]–[32]. Early in 1970s and late 1990s, mono-size small particles separation through a packed bed of mono-size large particles was studied by physical experiments [18], [39]–[48]. Later numerical models were also employed and the studies were extended to a much wider range of controlling variables on particle separation [3], [5], [9], [11], [15], [16], [19]–[21], [25]–[27], [29], [36], [37].

From the percentage recovery point of view, a jig can be classified as a complex system with multivariable nonlinear dynamics, large uncertainty according to external disturbances and both model structure and parameters, and multiple space and time scales dynamics [4]. Therefore, controlling recovery in jigs is, in general, not a straightforward task. These issues motivated the study of recovery behavior of jig by means of modeling and simulation tools. As pointed out by Dong *et al.* [4], there are several benefits to modeling jig: plant design and optimization, experimental design, testing research hypotheses, design and evaluation of control strategies, forecasting, analysis of plant-wide performance, and education [4], [28].

For design and optimization purposes of jig, the PET (potential energy theory) [8], [47] and DEM (discrete element method) [25], [38] simulation models are widely used to try understand the gravimetric concentration process in jigs. However, the above contributions don't address the recovery to the concentration of mineral particles widely distributed in size and density. Therefore, the investigations of Ospina and Usuga [15] and Ospina *et al.* [16] were the starting point. In these previous works, the authors evaluate the sensitivity of the mineral recovery through the variation of the operational parameters of the equipment, from a statistical and numerical analysis, using descriptive models, in the present investigation it is intended to model the jig through predictive intelligent systems.

Intelligent systems are in many places, from vehicles to cell phones and even in some common household appliances such as refrigerators and microwave ovens. This is nothing more than a black box that includes input/output. Among the scopes that can be achieved with this type of systems are the following: System identification techniques applied in real processes, investigating the current methodologies that are being studied; implement parameter estimates by means of neural networks to different systems, taking into account the

variables to manipulate and the data collected; validate with experimental data in real time the results obtained with the neural network model against using a mathematical model to confirm if there is a significant decrease in error and general improvements in the automatic control of processes, and encourage research and learning of neural networks for use in different areas of knowledge.

Artificial neural networks arise within the field of artificial intelligence, simulating the behavior of a biological neural network, in order to solve complex problems that would be very difficult to solve using conventional algorithms [49]–[52]. There are different types of artificial neural networks, which are used for different applications depending on their development. These networks are widely used for tasks such as: data classification, pattern detection, obtaining models of the retina of the eye and brain function, probability assessment, optimization, computer vision, and in the case of this research it was used for the prediction of variables in the mining industry.

However, in the field of modeling systems for mineral processing (grinding, classifying and concentrating), artificial neural networks are relatively recent, but their use is increasing in this type of systems due to the efficiency of the results that they can generate, avoiding the implementation of complex calculations with better performance [16], [24]. Currently, the intelligence systems by means of artificial neural networks can be summarized into four structures: i) supervised learning: the neural network learns a set of inputs and the desired outputs to solve the problem [53]–[56], ii) direct inverse learning: the neural network learns from the feedback of a system, so that, when the signal is obtained, it determines the parameters to be performed [52], [57]–[60], iii) utility backpropagation: this structure optimizes the mathematical equation that represents the system, where its main disadvantage is that it requires a model of the system to be analyzed [61]–[65] and iv) adaptive critical learning: similar to the utility backpropagation structure, but without the need for a model of the plant [66]–[68]. Although this type of structures are present and well accepted in different industrial processes, it is evident that in mineral processing applications and especially in the prediction of variables of interest such as mineral recovery, the existing studies of this type of design are based on simulations, this research being a starting point for the implementation of intelligent systems in gravimetric concentration equipment where experimental data obtained from a pilot scale jig is worked on.

This paper aims to use an ANN model to study the recovery percentage (yield) in a pilot jig and its interaction with other internal process variables (pulsation amplitude and frequency, water flow, particle size distribution and height of the artificial porous bed.). First an ANN model is obtained to predict recovery percentage by experimental data from the pilot jig. Then, the model is tested and assessed in terms of their ability to predict recovery with 65 other experimental tests different from those used for training the neural network.

### III. METHODOLOGY

Mineral processing is considered fundamental to the mining industry. Classically, the term mineral processing or

mineralurgy is used to describe the transformation operations involved in the upgrading and recovery of minerals [69]. These operations are carried out sequentially to obtain a raw material useful in subsequent processes or a final product desirable in the market. The operations that are grouped under the name mineral processing can be divided into four groups: size reduction, classification, concentration, and refining. Each stands out within a mineralurgical process, according to the mineralogical characteristics of the feed and the specifications of the final product. In a gravimetric concentration equipment a stream called feed is divided into two: a stream called concentrate, which has a high content of the species of interest and another stream called tails, in which this content is substantially decreased [2], [19]–[21]. Different operational parameters such as amplitude and frequency of water pulsation, bed thickness and feed flow characteristics affect the stratification process. In the following, the methodological application to obtain an artificial neural network model of a pilot-scale jig is shown.

The following methodology was proposed and followed step by step. This methodology attempts to bring together the theoretical component, the practical component, and the planning of the work.

- 1) Approach of an Experimental Design. Development of the Experimental Design. Assembly of jig at laboratory scale.
- 2) According to the proposed experiment design, the procedures to carry out the experimental tests on the laboratory scale jig and the data collection are planned.
- 3) Conducting tests in the jig at laboratory scale with alluvial ore sample suspensions to observe the concentration process in jigs, varying the proposed conditions. Sample collection and characterization.
- 4) Analysis of the results of the tests carried out.
- 5) Formulation of the ANN model that better predict the recovery percentage in jigs.
- 6) Validation of the obtained model with jig data at laboratory scale. The model was simulated in MATLAB®9.11(R2021b) using own code installed in a computer with an 8-core processor and 12GB of RAM. We used 216 samples (with sampling time  $T_s = 0.1$  s) for model identification and validation.

The aim is to estimate the percentage recovery of high-density minerals from low-density minerals in a jig according to a stratification of the particles present in a feed stream ( $F_a$ ) whose solids load is less than 10% in volume. The stratification is produced by the transmission of mechanical energy which is generated by the movement of a plunger that exerts pressure on the water ( $F_{h_2o}$ ) in the internal chamber of the jig in a harmonic way, generating a movement in pulses (ascent and descent) of the particulate system that enters the separation chamber of the jig, so that a stratification of the bed formed by the particles is obtained, which is later used to produce the separation of the minerals. The separation chamber of the jig is open to the atmosphere. Inside it there is a screen where a bed of particles is deposited with an intermediate density with respect to the minerals to be separated. The particle bed has an initial height  $H_0$  (packed bed) that rises to a height  $H_{max}$ , (fully fluidized bed)



according to the upward and downward movement of  $F_{h_{2o}}$ . The  $F_{h_{2o}}$  current contains water at a flow rate greater than or equal to the minimum fluidization velocity of the particles to be separated. The anharmonic motion of  $F_{h_{2o}}$  generates a hydrodynamic interaction between the two phases present in the process (solid-solid, solid-liquid interaction). This interaction alters the movement of the mineral particles in the separation chamber of the jig. The upward movement of water and mineral particles is called the fluidization stage. In this stage the mineral particles rise from a height  $H_0$  to a height  $H_{max}$ , initiating the stratification of the particles. At the beginning of stratification, the mineral particles with higher density and larger size tend to be deposited in the lower part of the bed, while the particles with lower density and smaller size are in the upper part of the bed. When the descent stage begins, the denser particles have a higher sedimentation velocity than the less dense particles, this allows that before the compaction of the bed, the heavier mineral particles are deposited quickly below the screen, obtaining after several cycles of pulsation, a complete separation of the mineral particles in two streams:  $F_{rejection}$  and  $F_{concentrated}$ . The process is carried out under ambient temperature conditions and no chemical reaction is present. A diagram of the general process is shown in Fig. 1. The jig from which actual data were taken to train and validate the model has the conditions reported in Table I.

A full factorial experimental design was developed involving the variables that exert the greatest control in the operation of the equipment (water flow, pulse amplitude, pulse frequency, granulometry, APB height) by means of an experimental matrix. This factorial design consists of five factors where the frequency, water flow and granulometry each have three levels (high, medium, low) and the amplitude and APB height each have two levels (high and low) resulting in a total of 108 tests plus replicates, a total of 216 tests. Table II summarizes the values considered for the different operating parameters. The response variable is the main metallurgical index (Recovery percentage (%R)).

Regarding the method of data collection and analysis. Primary data were used, which were collected through each of the ore samples generated from the experimental design (see Table I) by quantifying the gold mineral content in each of the 151 samples for identification and the 65 samples for validation, through two methods known as fire assay and time sequence analysis [7]. The effect of the operational parameters (see Table 2) on %R could be predicted from a sequential order of recovery percentage values (trend at equal time intervals).

The neural network method was selected because of its great capacity to adapt to different types of problems, the previous experience with the use of neural networks [70]–[73], and the ease of implementation of this type of technique, in addition to being a technique with great potential that is causing a revolution by proving to be the future of technology.

An artificial neural network (ANN) is an automatic learning and processing paradigm inspired by the functioning of the human nervous system [58], [59], [65]–[67], [74]. A neural network is composed of a set of neurons interconnected by links, where each neuron takes as inputs the outputs of the preceding neurons, multiplies each of these inputs by a weight

and, by means of an activation function, calculates an output. This output is in turn the input of the neuron it precedes. The union of all these interconnected neurons the artificial neural network [50], [51], [54], [55].



Fig. 1. Jig Diagram.

TABLE I. PILOT SCALE JIG CHARACTERISTICS

Feature	value
Mineral Feed Flow ( $F_a$ )	200 g/min
Water flow ( $F_{h_{2o}}$ )	[1.5-2.5] gal/min
Mineral size (granulometry)	[125-850] $\mu\text{m}$
Artificial porous bed height (APB)	[2.5-3.75] cm
Pulse amplitude	[5-7] mm
Pulse frequency	[5-9] Hz
APB particle diameter	3 mm

TABLE II. OPERATING PARAMETERS AND RESPECTIVE OPERATING LEVELS

Levels	High	Medium	Low
Frequency (Hz)-F	9	7	5
Pulse amplitude (mm)-A	7	-	5
Granulometry (mesh series Tyler)-T	+50 (1)	-50 +100 (2)	-100 (3)
Water flow (gal/min)-H	2,5	2,0	1,5
Artificial porous bed height (cm)	3,75	-	2,5



The artificial neural network as well as biological networks learn by repetition, and the more data you must train and the more times you train the network the better results you will get [62], [63], [67]. Training an ANN is a process that modifies the value of the weights associated with each neuron, so that the ANN can generate an output from the data presented in the input [52]. The weights are really the way the neuron learns. These weights will be modified in a certain way to adapt the value of the output in such a way as to minimize its error with respect to the real result that the artificial neuron should produce [55], [75].

Based on the above arguments, the following questions arise for this methodological development: What data are relevant for the management of the problem to be addressed?, which variables are relevant to address and manage this problem?, where can the data be obtained?, how to prepare and encode the data?, what type of network should be chosen?, how many hidden layers and how many neurons are necessary to manage the possible solution to the proposed problem?, what learning rule to choose?, and what initialization is given to the weights?. These data will be acquired by means of experimental tests using the pilot plant of the jig. The data obtained from the experimental tests will be organized in a spreadsheet to be later entered into the software that will be used to code the neural network (Matlab@R2021b)). The network to be designed will be initially selected with a configuration of one hidden layer with 13 neurons in each layer, five inputs, one output and a learning coefficient of 0.3 and random weights. However, once implemented, several tests will be performed to determine if a change in any of the parameters is necessary to obtain better results. The structure intended to be implemented for this proposed model is shown in Fig. 2. This structure has the possibility of being changed if a better alternative is discovered in the future during the process of development, research, and implementation.

As regards the activation functions, several functions were tested, among them the logsig, the ReLu, the softplus, the hardlim and the tansig, choosing in the end the logsig since it is the function that reaches a low margin of error in the shortest time for this specific process with the input data obtained.

Finally, it is possible to observe that the neural network was developed with its own code (the Matlab library was not used) along with the database with which it would be trained. The Matlab program consists of three parts: i) Network configuration: This is the main part, where the neural network is configured and trained. This was done using arrays of cells, so that it was possible to store arrays of different sizes in a single variable. Each row of the network variable is a different type of data, such as the weights of each layer, biases, errors, and so on, ii) Feedforward: In this file the forward propagation stage of the network was performed, where all the elements of the database are passed through the neural network, obtaining an output, and iii) Backpropagation: The backpropagation algorithm implemented for this network was gradient descent, where the aim is to minimize the error by calculating the partial derivatives of the error or cost function (mean squared error in this case), in terms of the weights of each neuron, modifying them in order to reduce the error in the output.

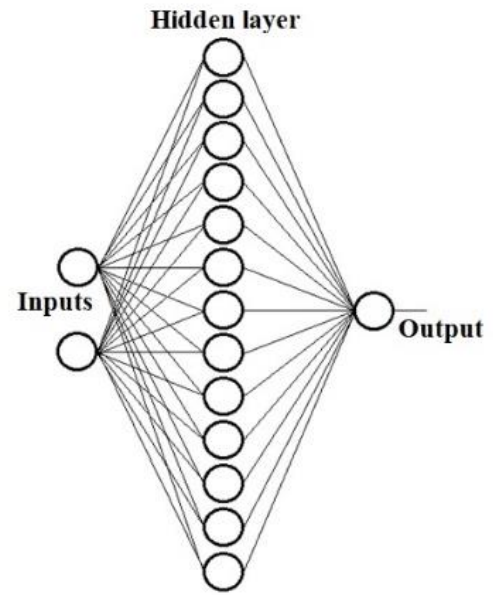


Fig. 2. Neural Network Configuration for Intelligent System.

Several tests were performed to verify the learning of this network and that the data delivered by it were consistent even with parameters that it had not received in the training stage, obtaining more than acceptable values, and thus proceeding to the implementation stage.

To assess the performance of ANN model, was followed the procedures suggested by Cruz et al.[58]. After visual evaluation and tests for fitting the behavior, the two criteria: Root Mean Square Error (RMSE), and coefficient of determination ( $R^2$ ) were applied [31], [32], [34] (see Eqs. (1) and (2)):

$$RSME = \sqrt{\frac{\sum_{t=1}^n (x - x_t)^2}{n}} \quad (1)$$

$$R^2 = 1 - \frac{\sum_{t=1}^n (x - x_t)^2}{\sum_{t=1}^n (x - \bar{x})^2} \quad (2)$$

where  $x$  is experimental data of the recovery percentage,  $x_t$  is ANN output,  $n$  is the sample size and  $\bar{x}$  is the mean of experimental data for the recovery percentage.

#### IV. RESULTS

This section presents the results obtained in this research, which includes the description of the intelligent implementation. The neural network was trained from plant experimental data, using the database "JigRNA.xlsx", which have 151 rows of data, the first column being the pulsation frequency value, the second the pulse amplitude value, the third the ore size, the fourth the water flow value, the fifth the APB value, and the sixth to the mineral recovery percentage value (output). Table III shows a portion of the recorded data.

Once the results of the experimental tests were obtained, the variables values were normalized in the interval 0 to 1, this in order to work with less uncertainty the data that would be entered into the neural network and have a free response of

engineering units. Eq (3) shows how the data were normalized and Table IV shows a portion of the normalized data.

$$Normalized_{value} = \frac{(value_{real}-value_{min})}{(value_{max}-value_{min})} \quad (3)$$

After normalization, a multilayer perceptron ANN with five inputs (corresponding to the five parameters used in Table IV) was created using proprietary code in a Matlab® script. The program identifies the five input and the output (%R). From the simulations performed, errors of less than 2% were obtained with 13 neurons in the hidden layer and starting the training process for 200 epochs. The results of the network without training and the training performance are shown in Fig. 3 and 4 respectively.

Fig. 3 shows the unsorted distribution of the data for the untrained neural network, once the training algorithm is started, it converges in only 40 epochs (see Fig. 4), making the network follow the input patterns that have been provided.

The performance of the ANN model was compared in two stages. In the first stage (Fig. 5), the identification data are shown with the absolute errors (maximum, average and minimum) that are reported in Table V.

TABLE III. RANDOM DATABASE USED TO TRAIN THE ANN

F	A	T	H	APB	%R
5	7	1	2	3,75	68,69
7	5	3	2,5	3,75	6,8
5	7	2	2,5	2,5	13,7
9	7	2	2	3,75	31,1
7	5	3	1,5	2,5	93,5
5	5	2	2	3,75	14,64
5	7	3	1,5	3,75	50,08

TABLE IV. NORMALIZED DATABASE USED TO TRAIN THE ANN

F	A	T	H	APB	%R
0	1	0	0,5	0	0,6905
0,5	0	1	1	1	0,0683
0	1	0,5	1	0,5451	0,1376
1	1	0,5	0,5	1	0,3126
0,5	0	1	0	0,545	0,9399
0	0	0,5	0,5	1	0,1471
0	1	1	0	1	0,5034

TABLE V. ERRORS IN IDENTIFICATION

Model	Max. Error	Mean Error	Min. Error
ANN	0,4449	0,0630	3,4382e-06

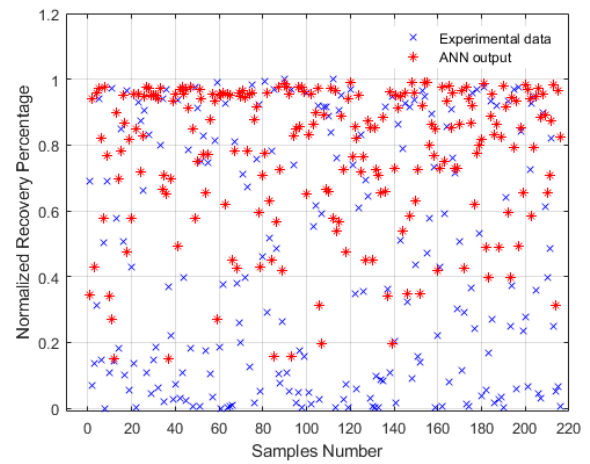


Fig. 3. ANN Response without Learning.

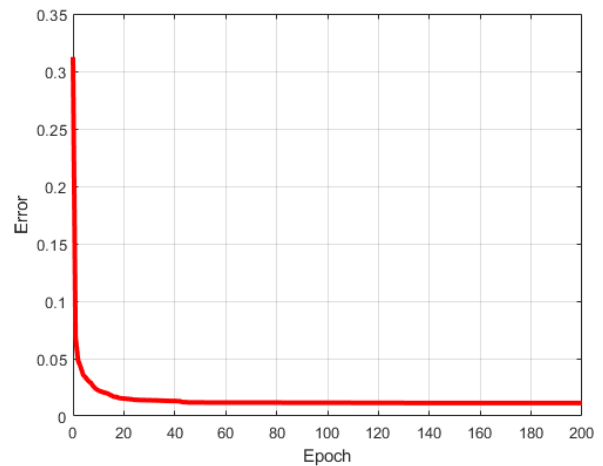


Fig. 4. Training Performance.

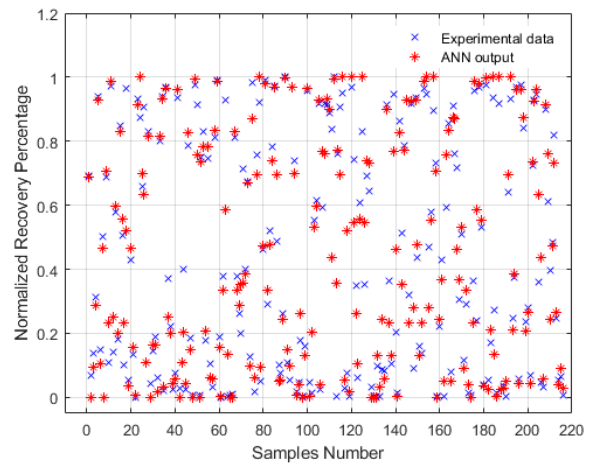


Fig. 5. ANN Performance.

In the second stage the model is compared by means of the validation data (Fig. 6) with their respective R<sup>2</sup> and RSME presented in Table VI. Additionally, Fig. 7 shows the histogram of the errors made by the Neural Network for the validation data.

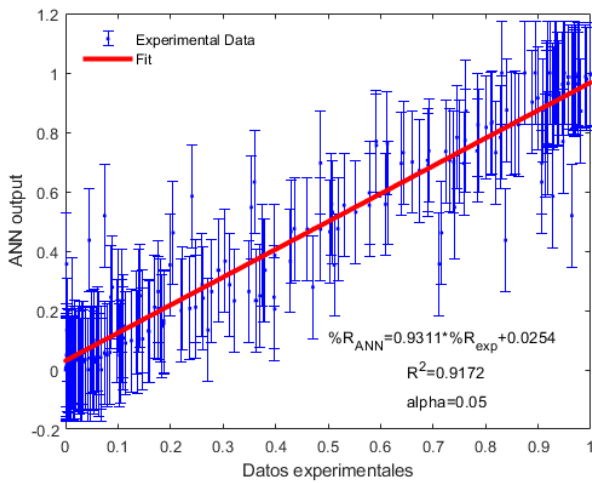


Fig. 6. Relationship of the Validation Data with the ANN Output.

TABLE VI. ERRORS IN VALIDATION

Model	R <sup>2</sup>	RSME
ANN	0,9172	0,105

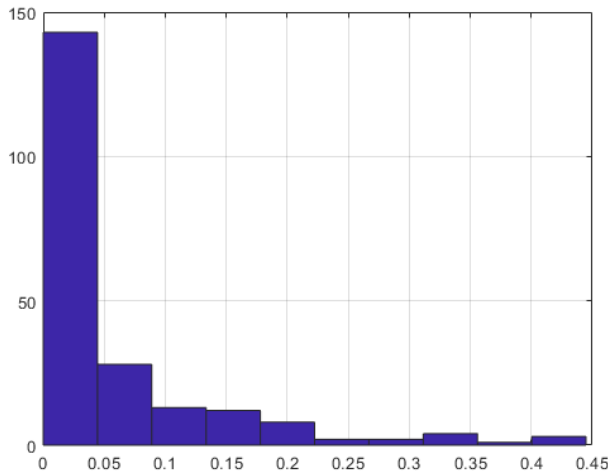


Fig. 7. Histogram of ANN Errors with Validation Data.

It can be seen from Fig. 5 to 7 and Tables V and VI that the performance of ANN model may be suitable for prediction of recovery percentage in gravimetric concentration equipment in the mineral processing industry, both in identification and validation, yielding errors of less than 2%, which is adequate for engineering purposes, especially to implement an optimization strategy for the jig concentration process.

The ANN model result in its matrix form can be expressed according to Eqs. (4) to (9). In the hidden layer Eq. (6) the inputs to the network  $X_{5 \times 1}$  were shown one by one (see Table 4). These result in an output  $Y_{13 \times 1}^1$  (Eq. (7)) for each neuron from applying the activation function  $f^1(a_{13 \times 1}^1)$  (logsig). Subsequently, the outputs of the neurons of the hidden layer are fed to the neuron of the output layer (Eq (8)) to obtain the output of the whole network or each input supplied (Eq. (9)).

$$a_{n \times 1}^k = W_{n \times p}^k \cdot Y_{p \times 1}^{k-1} + b_{n \times 1}^k \quad (4)$$

$$Y_{n \times 1}^k = f^k(a_{n \times 1}^k) \quad (5)$$

$$a_{13 \times 1}^1 = W_{13 \times 5}^1 \cdot X_{13 \times 1} + b_{13 \times 1}^1 \quad (6)$$

$$Y_{13 \times 1}^1 = f^1(a_{13 \times 1}^1) \quad (7)$$

$$a_{1 \times 13}^2 = W_{1 \times 13}^2 \cdot Y_{13 \times 1}^1 + b_{1 \times 1}^2 \quad (8)$$

$$Y_{1 \times 1}^2 = f^2(a_{1 \times 1}^2) \quad (9)$$

In Eq. (4) to (9), the  $W^k$  are the weights of each of the connections of the neurons with the upstream and downstream layers and the  $b^k$  are the biases of each of the neurons. These two parameters are trained for each input supplied to the network and are updated epoch by epoch until the neuron output is as close as possible to the experimental data. The values of these parameters after training and validation are shown in Tables VII to IX.

TABLE VII. INPUT-NEURON WEIGHT MATRIX HIDDEN LAYER ( $W_{13 \times 5}^1$ )

Neuron	F	A	T	H	APB
1	13,169	-8,118	9,593	-14,15	-7,127
2	13,338	23,438	14,351	6,829	-39,237
3	-2,497	-8,079	2,932	-1,547	-8,782
4	25,019	2,770	-40,937	3,161	2,032
5	2,287	5,335	1,892	3,824	3,860
6	-21,545	7,757	-12,758	-5,941	-14,059
7	-12,903	11,024	-9,002	0,628	8,4
8	9,771	-0,809	-20,987	-0,855	-1,302
9	13,742	-8,853	-6,582	-6,028	19,347
10	0,271	3,452	1,864	3,541	0,165
11	-7,925	5,39	2,966	7,421	-1,126
12	41,617	24,659	-23,697	-48,221	71,691
13	13,914	8,624	3,308	7,510	13,214

TABLE VIII. WEIGHT MATRIX NEURONS HIDDEN LAYER-NEURON OUTPUT LAYER ( $W_{1 \times 13}^2$ )

Neuron Hidden Layer	Output Neuron
1	6,5676
2	-4,5756
3	-6,4016
4	-12,4363
5	-20,2859
6	6,5559
7	-5,5759
8	15,4027
9	-10,4933
10	18,7184
11	-6,7827
12	5,4746
13	4,9074

In Table VII, each row represents a neuron of the hidden layer, and each column represents the connection to each of the inputs (corresponding to the five operational variables of the jig). Similarly in Table VIII, each row corresponds to an output of the hidden layer neurons, and each column is the connection from the hidden layer to the output neuron of the network. Finally, Table IX can be interpreted as follows: each row corresponds to the bias of each of the hidden layer neurons. The bias value for the output layer neuron was  $b_{1 \times 1}^2 = 6,4373$ .

TABLE IX. BIAS MATRIX NEURONS HIDDEN LAYER

Neuron Hidden Layer	Bias ( $b_{1 \times 1}^1$ )
1	-11,6266
2	-4,9026
3	11,9266
4	-0,1623
5	-7,4530
6	23,4868
7	-4,3211
8	4,6968
9	0,7725
10	-3,1677
11	-6,9047
12	-16,4843
13	-20,1379

## V. DISCUSSION OF RESULTS

In this paper we proposed as a contribution an ANN model for the estimation of the recovery percentage in a gold gravimetric concentration equipment known as a jig, which was based on the development of its own code in Matlab®, based on experimental data from a laboratory-scale jig (see Figure 1). With respect to traditional empirical and DEM-CFD modelling [15], [16], [20], [25], [26], artificial intelligence applications can become more efficient than the conventional techniques still used in the mineral processing industry as the ones proposed in this research and supported by [24] where is emphasized that in the ANN model, the input variables can include other controlling variables, such as other particle properties and system dimensions, feature that is very similar to phenomenologically based models.

Comparing both the identification and validation errors. It can be said that the performance of the ANN model adequately predicts the response variable (RSME=0.105), despite the great dependence and sensitivity of the recovery percentage with respect to variability. Of the five main operational parameters of the equipment (frequency and amplitude of pulsation, water flow, height of the artificial porous bed, and size of the mineral to be separated). The error in the ANN model could be decreased by two orders of magnitude if more delve into the number of neurons and hidden layers in the network that give a better fit with respect to the data provided. The above was evidenced in [58] where it is ratified that the use of different neural networks, whether they use the gradient method or the

convolutional ones, converges to a global minimum of the error.

The prediction performance of the model is compared by experimental data (see Fig. 5 and 6). Note that, since the validation data were recorded from laboratory-scale jig, It is necessary to be able to validate the model taking into account more ore samples and on a full-scale jig, as there could be a significant lag between the result of the pilot plant and the data recorded on an industrial-scale jig [3]. This type of inconvenience can be easily compensated with the application of scaling techniques, integrating the total plant and control design [3], [32].

The average prediction errors that were estimated using the validation data sequence for the model, are shown in Tables V and VI. We see that the ANN model fits well, but one of the limitations of these prediction errors is that a significant number of dynamics (Particle-particle interaction and particle-fluid interaction) remains unmodeled. This may be since when employed models that depend heavily on data (empirical), built through experimentation and observation, the interpretability of many other phenomena that occur in this type of process is limited to the little knowledge that this type of structures present, in addition to other effects that can only be adequately modeled in a phenomenological mathematical structure.

This paper has shown that the modeling framework based on ANN models may give models that are as useful, accurate, and reliable as with phenomenological modeling, even if the system is well understood. Such models may serve as an alternative that may be attractive especially for systems that are not well understood, as in the case of jig. Moreover, we believe ANN model developed in this framework has significant advantages over many other non-linear empirical modeling frameworks. The reason is that it admits interpretability of the model through the intuitive and easily understandable operating regime concept, and the fact that the machine learning models can be interpreted independently.

## VI. CONCLUSIONS

In this work, ANN model was developed, and their performance was evaluated by means of absolute errors. It can be said that the application of this type of model in real mineral processes can guarantee an adequate optimization of highly dynamic processes, since the expected results can be obtained in a shorter time without the need to use complex mathematical models, which are often difficult to obtain, or the phenomena involved in most of them cannot be understood in depth.

ANN models have a wide use in present day engineering. In this contribution, special attention is paid to control oriented models. As is well known, chemical, biochemical, and mining-metallurgical models are complex and nonlinear due to the multiple interacting phenomena, making them hard to implement in process control tasks. Although an ANN model (a type of black-box model) is not able to capture and predict the essential phenomena (mass, energy, and momentum transfer), gives significant rapid response with respect to the variables that are intended to intervene, thus providing strategies to rationally optimize and to control industrial scale

jigs since in this type of process a large amount of data can be obtained from both input and output variables.

With the implementation of this type of advanced modeling strategy, there was a significant reduction in the error when comparing the conventional empirical data against the experimental data using neural networks and, in turn, a better response to the operational variability of the processes was evidenced. Considering the results obtained, it can be affirmed that neural networks can be the pillar of the so-called fourth industrial revolution, proving to be useful in multiple fields, offering high efficiency and reliability in the optimization of industrial processes.

It is concluded from the identification and validation performed that the ANN model is very sensitive to the data provided. Further simulations on the ANN model, changing the number of neurons in the hidden layer could show very significant changes in the errors of both the identification and validation data.

As a future work derived from the present research, it is intended to complement the architecture of the proposed system by including other artificial intelligence models for the analysis of recovery percentage, using, for example, fuzzy logic, in order to enable the implementation of early warning systems and particle-particle interaction and particle-fluid interaction for monitoring recovery percentage.

#### ACKNOWLEDGMENTS

The authors would like to thank the Universidad de Cartagena – Colombia, the Instituto Tecnológico Pascual Bravo - Colombia for their support in the development of this research through the research project with ID: IN202204, and the Instituto Tecnológico Metropolitano - Colombia for their support through the research project entitled "Diseño y modelamiento de un concentrador gravimétrico con campo eléctrico para la recuperación de minerales" with ID: P20221.

#### REFERENCES

- [1] W. M. Ambrós, "Novos Aspectos da Estratificação de Partículas em Jigues Descontínuos," 2017.
- [2] W. M. Ambrós, "Jigging: A review of fundamentals and future directions," *Minerals*, vol. 10, no. 11, pp. 1–29, 2020, doi: 10.3390/min10110998.
- [3] E. F. Crespo, "Modeling segregation and dispersion in jigging beds in terms of the bed porosity distribution," *Miner. Eng.*, vol. 85, pp. 38–48, 2016, doi: 10.1016/j.mineng.2015.10.012.
- [4] K. J. Dong, S. B. Kuang, a. Vince, T. Hughes, and a. B. Yu, "Numerical simulation of the in-line pressure jig unit in coal preparation," *Miner. Eng.*, vol. 23, no. 4, pp. 301–312, Mar. 2010, doi: 10.1016/j.mineng.2009.10.009.
- [5] N. F. Feil, C. H. Sampaio, and H. Wotruba, "Influence of jig frequency on the separation of coal from the Bonito seam - Santa Catarina, Brazil," *Fuel Process. Technol.*, vol. 96, pp. 22–26, 2012, doi: 10.1016/j.fuproc.2011.12.010.
- [6] R. P. King, "Gravity separation," *Model. Simul. Miner. Process. Syst.*, pp. 233–267, 2001, doi: 10.1016/b978-0-08-051184-9.50011-0.
- [7] S. Kumar and R. Venugopal, "Performance analysis of jig for coal cleaning using 3D response surface methodology," *Int. J. Min. Sci. Technol.*, vol. 27, no. 2, pp. 333–337, 2017, doi: 10.1016/j.ijmst.2017.01.002.
- [8] F. W. Mayer, "Fundamentals of a potential theory of jigging process," in *7th International Mineral Processing Congress*, 1964, pp. 75–86.
- [9] B. K. Mishra and S. P. Mehrotra, "A jig model based on the discrete element method and its experimental validation," *Int. J. Miner. Process.*, vol. 63, no. 4, pp. 177–189, 2001, doi: 10.1016/S0301-7516(01)00053-9.
- [10] A. K. Mukherjee and B. K. Mishra, "An integral assessment of the role of critical process parameters on jigging," *Int. J. Miner. Process.*, vol. 81, no. 3, pp. 187–200, 2006, doi: 10.1016/j.minpro.2006.08.005.
- [11] M. A. Ospina, "Modelamiento de la hidrodinámica de la separación gravimétrica de minerales en jigs," *Universidad Nacional de Colombia*, 2014. [Online]. Available: <http://bdigital.unal.edu.co/50552/1/71265598.2015.pdf>
- [12] M. A. Ospina and M. O. Bustamante, "Hydrodynamic study of gravity concentration devices type JIG," *Rev. Prospect.*, vol. 13, no. 1, pp. 52–58, 2015, doi: <http://dx.doi.org/10.15665/rp.v13i1.359>.
- [13] W. M. Ambrós, C. H. Sampaio, B. G. Cazacliu, P. N. Conceição, and G. S. dos Reis, "Some observations on the influence of particle size and size distribution on stratification in pneumatic jigs," *Powder Technol.*, vol. 342, pp. 594–606, 2019, doi: 10.1016/j.powtec.2018.10.029.
- [14] M. A. Ospina, A. B. Barrientos, and M. O. Bustamante, "Influence of the pulse wave in the stratification of high density particles in a JIG device," *Rev. Tecnológicas*, vol. 19, no. 36, pp. 13–25, 2016, doi: 10.22430/22565337.585.
- [15] M. A. Ospina and L. M. Usuga, "Effect of Hydrodynamic forces on mineral particles trajectories in Gravimetric concentrator Type Jig," *Politecnica*, vol. 14, no. 27, pp. 68–79, 2018, doi: 10.33571/rpolitec.v14n27a7.
- [16] M. A. Ospina, L. M. Usuga, G. E. Chanchí, and S. Gómez, "Assessment and modeling of electric force in a jig device," *ARNP J. Eng. Appl. Sci.*, vol. 16, no. 23, pp. 2581–2588, 2021, [Online]. Available: [http://www.arnpjournals.org/jeas/research\\_papers/rp\\_2021/jeas\\_1221\\_8778.pdf](http://www.arnpjournals.org/jeas/research_papers/rp_2021/jeas_1221_8778.pdf)
- [17] F. Pita and A. Castilho, "Influence of shape and size of the particles on jigging separation of plastics mixture," *Waste Manag.*, vol. 48, pp. 89–94, 2016, doi: 10.1016/j.wasman.2015.10.034.
- [18] R. X. Rong and G. J. Lyman, "A new energy dissipation theory of jig bed stratification . Part 1 " energy dissipation analysis in a pilot scale baum jig," *Int. J. Miner. Process.*, vol. 37, pp. 165–188, 1993, doi: 10.1016/0301-7516(93)90025-6.
- [19] S. M. Viduka, Y. Q. Feng, K. Hapgood, and M. P. Schwarz, "Discrete particle simulation of solid separation in a jigging device," *Int. J. Miner. Process.*, vol. 123, pp. 108–119, 2013, doi: 10.1016/j.minpro.2013.05.001.
- [20] S. . Viduka, Y. . Feng, K. Hapgood, and P. Schwarz, "CFD-DEM investigation of particle separations using a sinusoidal jigging profile," *Adv. Powder Technol.*, vol. 24, no. 2, pp. 473–481, 2013, doi: 10.1016/j.apt.2012.11.012.
- [21] S. Viduka, Y. Feng, K. Hapgood, and P. Schwarz, "CFD-DEM investigation of particle separations using a Trapezoidal jigging profile," in *Ninth International Conference on CFD in the Minerals and Process Industries*, 2012, vol. 24, no. 2, pp. 473–481.
- [22] L. C. Woollacott and M. Silwamba, "An experimental study of size segregation in a batch jig," *Miner. Eng.*, vol. 94, pp. 41–50, 2016, doi: 10.1016/j.mineng.2016.04.003.
- [23] L. C. Woollacott, "The impact of size segregation on packing density in jig beds: An X-ray tomographic study," *Miner. Eng.*, vol. 131, no. June 2018, pp. 98–110, 2019, doi: 10.1016/j.mineng.2018.10.017.
- [24] S. M. Arifuzzaman, K. Dong, H. Zhu, and Q. Zeng, "DEM study and machine learning model of particle percolation under vibration," *Adv. Powder Technol.*, vol. 33, no. 5, p. 103551, 2022, doi: 10.1016/j.apt.2022.103551.
- [25] Y. K. Xia, F. F. Peng, and E. Wolfe, "CFD simulation of fine coal segregation and stratification in jigs," *Int. J. Miner. Process.*, vol. 82, no. 3, pp. 164–176, 2007, doi: 10.1016/j.minpro.2006.10.004.
- [26] Y. K. Xia and F. F. Peng, "Numerical simulation of behavior of fine coal in oscillating flows," *Miner. Eng.*, vol. 20, pp. 113–123, 2007, doi: 10.1016/j.mineng.2006.06.004.
- [27] K. Asakura, M. Mizuno, M. Nagao, and S. Harada, "Numerical Simulation of Particle Motion in a Jig Separator," in *5th Join ASME JSME Fluids Engineering Conference*, 2007, pp. 385–391.

- [28] M. A. A. Aziz, K. M. Isa, N. J. Miles, and R. A. Rashid, "Pneumatic jig: effect of airflow, time and pulse rates on solid particle separation," *Int. J. Environ. Sci. Technol.*, no. January, pp. 1–12, 2018, doi: 10.1007/s13762-018-1648-4.
- [29] L. Chica Osório, M. Ospina Alarcón, and M. Bustamante Rúa, "Uso de CFD para la simulación de procesos mineralúrgicos de concentración gravimétrica," *Uso CFD para la simulación procesos Miner. Conc. gravimétrica*, vol. 10, no. 1, pp. 85–96, 2012.
- [30] S. Cierpisz, "A dynamic model of coal products discharge in a jig," *Miner. Eng.*, vol. 105, pp. 1–6, 2017, doi: 10.1016/j.mineng.2016.12.010.
- [31] S. Cierpisz, M. Kryca, and W. Sobierajski, "Control of coal separation in a jig using a radiometric meter," *Miner. Eng.*, vol. 95, pp. 59–65, 2016, doi: 10.1016/j.mineng.2016.06.014.
- [32] S. Cierpisz and J. Joostberens, "Monitoring of coal separation in a jig using a radiometric density meter," *Meas. J. Int. Meas. Confed.*, vol. 88, pp. 147–152, 2016, doi: 10.1016/j.measurement.2016.03.060.
- [33] S. Cierpisz, "Some Aspects of the Optimal Control of the Coal Separation Process," *IFAC Proc. Ser.*, vol. 16, no. 15, pp. 445–455, 1984, doi: 10.1016/s1474-6670(17)64298-8.
- [34] S. Cierpisz and E. Jachnik, "On Optimal Control of Coal Separation Process," *IFAC Proc. Vol.*, vol. 20, no. 8, pp. 123–127, 1987, doi: 10.1016/s1474-6670(17)59081-3.
- [35] Y. Pan and M. Dagnew, "A new approach to estimating oxygen off-gas fraction and dynamic alpha factor in aeration systems using hybrid machine learning and mechanistic models," *J. Water Process Eng.*, vol. 48, no. May, 2022, doi: 10.1016/j.jwpe.2022.102924.
- [36] C. R. . Abreu, F. W. Tavares, and M. Castier, "Influence of particle shape on the packing and on the segregation of spherocylinders via Monte Carlo simulations," *Powder Technol.*, vol. 134, no. 1–2, pp. 167–180, 2003, doi: 10.1016/S0032-5910(03)00151-7.
- [37] K. Asakura, M. Mizuno, M. Nagao, and S. Harada, "Numerical Simulation of Particle Motion in a Jig Separator," in *5th Join ASME JSME Fluids Engineering Conference*, 2007, pp. 385–391.
- [38] A. J. . Beck and P. N. Holtham, "Computer simulation of particle stratification in a two-dimensional batch jig," *Miner. Eng.*, vol. 6, no. 5, pp. 523–532, 1993.
- [39] Y. Kawashima, J. Yassuke, and H. Isao, "Mechanical consideration of pneumatic jig vibration mechanism: 2nd report, Bed damping action," *Proc. Japanese Soc. Mech.*, vol. 40, no. 333, pp. 1309–1317, 1974, doi: 10.1299/kikai1938.40.1309.
- [40] H. J. Witteveen, "The Response of Uniform Jig Bed in Terms of the Porosity Distribution," 1995.
- [41] Y. Kawashima and J. Yasushikai, "Mechanical consideration of pneumatic jig vibration mechanism: 1st report, Vibration characteristics in the case of no damping," *Proc. Japanese Soc. Mech.*, vol. 39, no. 317, pp. 167–175, 1973.
- [42] R. . King, "A quantitative model for gravity separation unit operations that rely on stratification," *APCOM 87 Proc. 20th Int. ...*, vol. 2, pp. 141–151, 1987, [Online]. Available: <http://www.saimm.co.za/Conferences/Apcom87Metallurgy/141-King.pdf>
- [43] G. J. Lyman, "Review of Jigging Principles and Control," *Coal Prep.*, vol. 11, no. 3–4, pp. 145–165, 1992, doi: 10.1080/07349349208905213.
- [44] R. . Rong and G. . Lyman, "Modelling jig bed stratification in a pilot scale baum jig," *Miner. Eng.*, vol. 4, no. 5–6, pp. 611–623, 1991, doi: 10.1016/0892-6875(91)90007-I.
- [45] R. . Rong and G. . Lyman, "The effect of jigging time and air cycle on bed stratification in a pilot scale Baum jig," *Fuel*, vol. 71, no. 1, pp. 115–123, 1992, doi: 10.1016/0016-2361(92)90201-X.
- [46] R. . Rong and G. . Lyman, "The mechanical behavior of water and gas phases in a pilot scale baum jig," *Coal Prep.*, vol. 9, pp. 85–106, 1991, doi: 10.1080/07349349108960559.
- [47] L. M. Tavares and R. P. King, "A Useful Model for the Calculation of the Performance of Batch and Continuous Jigs," *Coal Prep.*, vol. 15, no. 3–4, pp. 99–128, 1995, doi: 10.1080/07349349508905291.
- [48] R. Srinivasan, B. K. Mishra, and S. P. Mehrotra, "Simulation of Particle Stratification in Jigs," *Coal Prep.*, vol. 20, no. 1–2, pp. 55–70, 1999, doi: 10.1080/07349349908945592.
- [49] L. Zhang, Y. Xue, Q. Xie, and Z. Ren, "Analysis and neural network prediction of combustion stability for industrial gases," *Fuel*, vol. 287, no. July 2020, p. 119507, 2021, doi: 10.1016/j.fuel.2020.119507.
- [50] Y. Oh, Y. Kim, K. Na, and B. D. Youn, "A deep transferable motion-adaptive fault detection method for industrial robots using a residual-convolutional neural network," *ISA Trans.*, no. xxxx, 2021, doi: 10.1016/j.isatra.2021.11.019.
- [51] M. Jalanko, Y. Sanchez, V. Mahalec, and P. Mhaskar, "Adaptive system identification of industrial ethylene splitter: A comparison of subspace identification and artificial neural networks," *Comput. Chem. Eng.*, vol. 147, p. 107240, 2021, doi: 10.1016/j.compchemeng.2021.107240.
- [52] S. Chen, J. Yu, and S. Wang, "One-dimensional convolutional neural network-based active feature extraction for fault detection and diagnosis of industrial processes and its understanding via visualization," *ISA Trans.*, no. xxxx, 2021, doi: 10.1016/j.isatra.2021.04.042.
- [53] X. Hu, G. Li, P. Niu, J. Wang, and L. Zha, "A generative adversarial neural network model for industrial boiler data repair," *Appl. Soft Comput.*, vol. 104, p. 107214, 2021, doi: 10.1016/j.asoc.2021.107214.
- [54] P. Kumar, J. B. Rawlings, and S. J. Wright, "Industrial, large-scale model predictive control with structured neural networks," *Comput. Chem. Eng.*, vol. 150, p. 107291, 2021, doi: 10.1016/j.compchemeng.2021.107291.
- [55] G. Gravanis, I. Dragogias, K. Papakiriakos, C. Ziogou, and K. Diamantaras, "Fault detection and diagnosis for non-linear processes empowered by dynamic neural networks," *Comput. Chem. Eng.*, vol. 156, p. 107531, 2022, doi: 10.1016/j.compchemeng.2021.107531.
- [56] S. K. Varanasi, A. Daemi, B. Huang, G. Slot, and P. Majoko, "Sparsity constrained wavelet neural networks for robust soft sensor design with application to the industrial KIVCET unit," *Comput. Chem. Eng.*, vol. 159, p. 107695, 2022, doi: 10.1016/j.compchemeng.2022.107695.
- [57] A. Brusaferrri, M. Matteucci, S. Spinelli, and A. Vitali, "Learning behavioral models by recurrent neural networks with discrete latent representations with application to a flexible industrial conveyor," *Comput. Ind.*, vol. 122, p. 103263, 2020, doi: 10.1016/j.compind.2020.103263.
- [58] Y. J. Cruz, M. Rivas, R. Quiza, A. Villalonga, R. E. Haber, and G. Beruvides, "Ensemble of convolutional neural networks based on an evolutionary algorithm applied to an industrial welding process," *Comput. Ind.*, vol. 133, 2021, doi: 10.1016/j.compind.2021.103530.
- [59] Y. Wu and L. Huang, "An intelligent method of data integrity detection based on multi-modality fusion convolutional neural network in industrial control network," *Meas. J. Int. Meas. Confed.*, vol. 175, no. January, p. 109013, 2021, doi: 10.1016/j.measurement.2021.109013.
- [60] M. S. Alhajerri, J. Luo, Z. Wu, F. Albalawi, and P. D. Christofides, "Process structure-based recurrent neural network modeling for predictive control: A comparative study," *Chem. Eng. Res. Des.*, 2022, doi: 10.1016/j.cherd.2021.12.046.
- [61] W. J. Alvarenga et al., "Online learning of neural networks using random projections and sliding window: A case study of a real industrial process," *Eng. Appl. Artif. Intell.*, vol. 100, no. October 2020, p. 104181, 2021, doi: 10.1016/j.engappai.2021.104181.
- [62] J. Krishnaiah, C. S. Kumar, and M. A. Faruqi, "Modelling and control of chaotic processes through their bifurcation diagrams generated with the help of recurrent neural network models. Part 2: An industrial study," *J. Process Control*, vol. 16, no. 1, pp. 67–79, 2006, doi: 10.1016/j.jprocont.2005.04.003.
- [63] C. H. Lu and C. C. Tsai, "Generalized predictive control using recurrent fuzzy neural networks for industrial processes," *J. Process Control*, vol. 17, no. 1, pp. 83–92, 2007, doi: 10.1016/j.jprocont.2006.08.003.
- [64] R. Jia, S. Zhang, and F. You, "Nonlinear soft sensor development for industrial thickeners using domain transfer functional-link neural network," *Control Eng. Pract.*, vol. 113, no. May, p. 104853, 2021, doi: 10.1016/j.conengprac.2021.104853.
- [65] B. LI, W. TIAN, C. ZHANG, F. HUA, G. CUI, and Y. LI, "Positioning error compensation of an industrial robot using neural networks and experimental study," *Chinese J. Aeronaut.*, vol. 35, no. 2, pp. 346–360, 2022, doi: 10.1016/j.cja.2021.03.027.
- [66] I. Chakraborty, B. M. Kelley, and B. Gallagher, "Industrial control system device classification using network traffic features and neural

- network embeddings,” *Array*, vol. 12, p. 100081, 2021, doi: 10.1016/j.array.2021.100081.
- [67] T. Walser and A. Sauer, “Typical load profile-supported convolutional neural network for short-term load forecasting in the industrial sector,” *Energy AI*, vol. 5, p. 100104, 2021, doi: 10.1016/j.egyai.2021.100104.
- [68] M. Benne, B. Grondin- Perez, J. P. Chabriat, and P. Hervé, “Artificial neural networks for modelling and predictive control of an industrial evaporation process,” *J. Food Eng.*, vol. 46, no. 4, pp. 227–234, 2000, doi: 10.1016/S0260-8774(00)00055-8.
- [69] M. C. Fuerstenau and N. Kenneth, *Principles of Minerals Processing*. New York: SEM, 2003.
- [70] P. Nenninger and D. Streitferdt, “On the Importance of Tailorable Processes in the Development of Embedded Industrial Automation Systems,” vol. 41, no. 2. *IFAC*, 2008. doi: 10.3182/20080706-5-kr-1001.00382.
- [71] G. Bloch and T. Dencœur, “Neural networks for process control and optimization: Two industrial applications,” *REE, Rev. L’Electricite L’Electronique*, no. 7–8, pp. 31–41, 2001, doi: 10.3845/ree.2001.074.
- [72] C. Puebla, “Industrial process control of chemical reactions using spectroscopic data and neural networks: A computer simulation study,” *Chemom. Intell. Lab. Syst.*, vol. 26, no. 1, pp. 27–35, 1994, doi: 10.1016/0169-7439(94)90015-9.
- [73] A. E. Smith and C. H. Dagli, “Controlling industrial processes through supervised, feedforward neural networks,” *Comput. Ind. Eng.*, vol. 21, no. 1–4, pp. 247–251, 1991, doi: 10.1016/0360-8352(91)90096-O.
- [74] J. Polaczek and J. Sosnowski, “Exploring the software repositories of embedded systems: An industrial experience,” *Inf. Softw. Technol.*, vol. 131, no. August 2019, 2021, doi: 10.1016/j.infsof.2020.106489.
- [75] L. Chen, J. Cao, K. Wu, and Z. Zhang, “Application of Generalized Frequency Response Functions and Improved Convolutional Neural Network to Fault Diagnosis of Heavy-duty Industrial Robot,” *Robot. Comput. Integr. Manuf.*, vol. 73, no. August 2021, p. 102228, 2022, doi: 10.1016/j.rcim.2021.102228.



# Risk Prediction Applied to Global Software Development using Machine Learning Methods

Hossam Hassan<sup>1</sup>, Manal A. Abdel-Fattah<sup>2</sup>, Amr Ghoneim<sup>3</sup>  
Information Systems Department, Helwan University, Egypt<sup>1,2</sup>  
Computer Sciences Department, Helwan University, Egypt<sup>3</sup>

**Abstract**—Software companies aim to develop high-quality software projects with the best global resources at the best cost. To achieve this global software development (GSD), an approach should be used which adopts work on projects across multiple distributed locations, and this is also known as distributed development. When companies attempt to implement GSD, they face numerous challenges owing to the nature of GSD and its differences from traditional methods. The objectives of this study were to identify the top software development factors that affect the overall success or failure of a software project using exploratory data analysis to find relationships between these factors, and to develop and compare risk prediction models that use machine learning classification techniques such as logistic regression, decision tree, random forest, support vector machine, K-nearest neighbors, and Naive Bayes. The findings of this study are as follows: in GSD, the top 18 factors influencing the software project are listed; and experiments show that the logistic regression and random forest models provide the best results, with an accuracy of 89% and 85%, respectively, and an area under the curve of 73% and 71%, respectively.

**Keywords**—Global software development; distributed development; risk prediction model; machine learning

## I. INTRODUCTION

The entire software development approach has permanently changed in the last two decades to support distributed development environments with distributed teams [1]. This strategy can be described as a contract between two parties, with the client representing advanced countries and the vendor representing developing countries, with the goal of achieving mutual interests [2]. Therefore, the main reason for the widespread use of global software development (GSD) is that clients worldwide need highly specialized resources and tools at a reasonable price [3].

In addition, GSD has seen a considerable increase in contracts and business in recent years. The use of distributed development teams in various time zones and geographic locations may be referred to as the ‘new age’ of development projects employing GSD [4]. The affordable price of GSD is a significant factor contributing to its appeal. Consequently, there has been great success in the mutual benefit between clients and vendors [5], [6]. Some benefits of adopting GSD include sharing knowledge, using the most recent technologies, access to resources, economic benefits, lower expenses, and successful overall project completion [7], [8].

In addition, the challenges and limitations that have a significant effect on GSD should be pointed out. For example,

it can be difficult for distributed teams to communicate with each other and work together due to language barriers, cultural norms and limits, time zones, leadership, team capabilities, and project management [9]–[11]. One of the most serious issues confronting GSD is the location, distance, and communication between the distributed teams [12]. In addition, the problem of team communication has been solved owing to the benefits of using agile methods such as scrum [8].

However, risks remain when clients attempt to adopt and use this approach in their projects. It can also yield the opposite results if it is misused. In the beginning, the term "risk" can be identified as a collection of software project characteristics, situations, and regulations that present a hazard to a project's overall success. It is also important to determine how often these risks occur, and how to prepare for them [13].

The Project Management Institute (PMI) shows that most risk management methods and procedures are ignored and thrown out, especially in the IT industry, because they are too general or only work in a specific situation [14]. Despite this, software projects that use techniques and tools to predict risk can detect approximately 70% and avoid 90% of harmful risks [15].

So, companies need to know the benefits and the risks of adopting the GSD approach, in an early stage of the development, to avoid any financial loss. In addition, companies need to also know if adopting GSD approach is suitable for their project or it will have negative results. Therefore, a software risk prediction model using the machine learning classification techniques was provided in this study, to make a prediction of the success or failure of the software project in the domain of GSD.

In this study, the following are discussed: First, previous systematic literature reviews were reviewed to identify the top software risk factors affecting GSD. Second, a dataset was collected from software projects in various regions of the world. Third, exploratory data analysis (EDA) was conducted to find different insights and correlations between these factors and each other. Fourth, software risk prediction models were built using different supervised machine learning classification techniques. Finally, software risk prediction models were evaluated and represented to determine the best model suitable for the GSD approach.

As a result, this study answers two main questions in the section between parentheses. RQ1: Which software risk factors

are essential to the GSD domain and significantly affect the software risk prediction? (Section III-A)

RQ2: What are the best machine learning techniques for software risk prediction in GSD? (Section V)

The remainder of this paper is organized as follows. Section II presents related work. The methodology is described in detail in Section III. Section IV presents an examination and measurement of the precision. Section V presents the results of the proposed model. Section VI discusses the validity threats. In Section VII the conclusion is presented, finally, in Section VIII, additional work is listed to be considered in the future.

## II. RELATED WORKS

This section presents two types of studies. The first one concerns a systematic literature review related to GSD factors, and the second one is related to software risk prediction using machine learning and other techniques.

### A. Systematic Literature Review for GSD Factors:

In [5], an empirical investigation was conducted to figure out the top requirements of engineering (RE) practices in GSD. Among the 66 practices, the results showed that only six key factors play an important role in GSD, as listed below:

- 1) Identify and consult with system stakeholders.
- 2) Prioritize requirements.
- 3) Define system boundaries.
- 4) Define standard templates for requirements.
- 5) Check if requirements document meet your standards.
- 6) Uniquely identify each requirement.

The dataset was collected by conducting an online survey questionnaire. For the evaluation of these factors, 56 experts from GSD were involved. Limitation and future work: the questionnaire relied only on closed questions and focused only on the company size, testing these factors, and trying to develop a framework to be used in the future.

In [16], the authors tried to prioritize the success factors that affect requirement change management (RCM) in the GSD. Fuzzy logic analytical hierarchy progress (FAHB) was used to conduct the prioritization. The result of this study was to find out the RCM success factors and categorize them into four groups: team, technology, process, and organizational management. The dataset was collected by conducting a questionnaire survey and retrieved around 81 responses. Evaluation metrics for the prioritization were conducted by using experts' responses. Limitations and future work: sample size of the dataset needs to be widened, and organization size and types should be considered, in addition, success factors, barriers, and best practices need more investigation and analysis.

The authors in [17], focused on scaling agile projects in the domain of GSD. They mapped 44 agile practices to the SAFe Framework. Instructions were given for how the SAFe practices can be used in agile global software development (AGSD) projects. The dataset was collected by reviewing 86 studies. Of these studies, only 24 papers discussed the scaling of agile, from which the authors selected 44 practices to be mapped on the SAFe Framework. Limitations and future work:

(AGSD) practices need to be evaluated and should also be tested in the real industry. In addition, the mapped process of these practices needs to be evaluated.

### B. Software Risk Prediction Models:

In [15], a software risk prediction model was created based on risk analysis of the project by using its context history and project characteristics in the software development life cycle (SDLC) as shown in Fig. 1. The model is called the Atropos model and consisted of six main phases listed below:

- 1) Data Gathering through interface and bulk uploading.
- 2) Similarity by characteristics of the project.
- 3) Store context histories of the project.
- 4) Similarities by context histories.
- 5) Recommendation of any potential risks.
- 6) Risk management and monitoring.

The dataset was collected based on 153 software projects from a financial company. Evaluation metrics of the model showed an acceptance rate of 73% and an accuracy rate of 83%, and these results were assessed by 18 experts. Limitations and Future work for the model are to improve the model's accuracy, to improve the proposed model and methodology, additional use of prototype, the number of practitioners, and the duration of the study (5 weeks only).

In [9], artificial neural network (ANN) model was created to predict the risk factors in GSD. The model used algorithms such as Levenberg–Marquardt, Bayesian Regularization, and Scaled Conjugate Gradient. The dataset was collected by sending 760 questionnaires to companies. 390 were received, and 116 were rejected, leaving 274 responses that were used as the primary data set. Evaluation metrics of the model were conducted by using least mean square error (MSE), and the results showed that Bayesian Regularization gave better results as compared with the other two approaches and matched the results from these studies [18], [19]. Limitations and Future work for the model are the sample dataset needs to include many companies and random data collection should be used to generalize the model, also the author recommended to use deep learning to get more insights and accurate results in the future.



Fig. 1. Shows the Atropos Six-Stage Model [15].

In [20], the authors provided a software reliability prediction algorithm. They used fuzzy logic and ANN in their model. The dataset was collected from John Musa of Bell Laboratories and received from the IEEE repository. Evaluation metrics of the model were conducted by using root mean square deviation error (RMSE) and showed that the fuzzy-neural method was the best compared to other algorithms. Limitations and future work for the model: the model is restricted to one factor (time to failure). In addition, many software risk factors should be used to evaluate this model better.

The authors in [21] developed a fuzzy logic hybridized framework for software risk prediction models during the decision-making process. Technique for order of preference by similarity to ideal solution (IF-TOPSIS), fuzzy decision-making trial and evaluation laboratory (DEMATEL), and crow search algorithm (CSA), optimized adaptive neuro-fuzzy inference system (ANFIS) were used for the software model prediction. The dataset consisted of 93 software projects, 70% used for training and the remaining used for testing and validating the model. The results showed that integrated fuzzy was accurate in software risk prediction. Limitations and future work: make a set of decisions and use many software factors and advanced machine learning techniques to improve and validate the results.

To reduce cost risks, the authors of [22] amplified the constructive cost model (COCOMO-II) in the GSD context. The dataset was collected by conducting a questionnaire and

receiving around 175 responses. Evaluation metrics of the model were conducted by using Magnitude of Relative Estimates (MRE) and experts' judgment. Limitations and Future Work: the model is in an early stage and needs more validation and evaluation. In addition, mathematical or machine learning (ML) techniques may be used in the future.

In [23], the authors developed ML models for defect prediction in the domain of software reliability and performance. The models were built using ANN, random forest (RF), random tree (RT), decision table (DT), linear regression (LR), Gaussian processes (GP), SMOreg, and M5P. The dataset for these models was from the NASA promise repository. The results showed that the combination of different ML algorithms is effective in the prediction of software defects. Evaluation matrices used were correlation coefficient ( $R^2$ ), mean absolute error (MAE), (RMSE), relative absolute error (RAE), and root relative squared error (RRSE). Limitation and Future works: different datasets and ML algorithms can be used to evaluate the results. In addition, more investigation into software factors should be conducted to improve these results.

Most previous studies concentrated on a limited number of factors, as summarized in Table I. In addition, the dataset needs to be enlarged to include more regions, and (ML) techniques need to be improved and evaluated using real data from software companies, as will be provided in the subsequent section.

TABLE I. OVERVIEW OF PREVIOUS RESEARCH STUDIES FOR SOFTWARE PREDICTION MODELS

Reference	Dataset	ML Techniques and algorithms	Evaluation metrics	Limitation and Future work
(Filippetto et al, 2021) [15]	The dataset was collected based on 153 software projects from a financial company.	Risk analysis of the project by using its context history and project characteristics in the (SDLC).	Acceptance rate of 73% and an Accuracy rate of 83%, and these results were assessed by experts	1. Improve the proposed model methodology and accuracy. 2. Additional use of prototype. 3. Number of practitioners and the duration of the case study should be increased.
(Ifikhar et al, 2021) [9]	The dataset was collected by sending 760 questionnaires to companies. 390 were received, and 116 were rejected, leaving 274 valid responses.	(ANN) model was created to predict the risk factors in GSD such as: Levenberg–Marquardt, Bayesian Regularization, and Scaled Conjugate Gradient.	MSE	1. the sample dataset needs to include many companies and random data collection should be used to generalize the model. 2. Deep learning should be used to get more accurate results.
(Sahu et al, 2018) [20]	The dataset was collected from John Musa of Bell Laboratories and received from the IEEE repository.	Fuzzy logic and ANN were used for building a software reliability prediction model.	RMSE	1. Model was restricted to one factor (time to failure). 2. Many software risk factors should be used to evaluate this model better.
(Suresh et al, 2021) [21]	The dataset consisted of 93 software projects, 70% used for training and the remaining used for testing and validating the model.	Fuzzy logic hybridized framework for software risk prediction models during the decision-making process.	CSA	1. Make a group of decisions making and use sophisticated ML techniques 2. Use many software factors
(Khan et al, 2021) [22]	The dataset was collected by conducting a questionnaire and receiving around 175 responses	Amplified COCOMO-II Model in the context of GSD.	(MRE, experts' judgment	1. The model is in an early stage and needs more validation. 2. Mathematical or ML techniques may be used in the future.
(Assim et al, 2020) [23]	The dataset for these models was from the NASA promise repository	ANN, RF, RT, DT, LR, GP, SMOreg, and M5P were used.	( $R^2$ ), (MAE), (RMSE), (RAE) and (RRSE)	1. different datasets and ML algorithms can be used to evaluate the results. 2. Investigation into more software factors to improve these results.

### III. RESEARCH METHODOLOGY

This section describes the methodology used to develop the GSD-applicable software risk-prediction model. Fig. 2 illustrates the six phases of the proposed model. Systematic Literature review (SLR) analysis (Section III-A), dataset selection (Section III-B), dataset preprocessing (Section III-C), modeling (Section III-D), experimental evaluation (Section IV), and risk prediction results (Section V).

#### A. Systematic Literature Review Analysis:

The proposal to build the software risk prediction model was based on many systematic literature reviews (SLR) that collected the software risk factors that affect the software in GSD. SLRs included empirical studies published between 2018 and 2022. After reviewing these studies, a list of 145 factors essential to software project success was created. (Available in Appendix "I").

Then, these factors were analyzed and reprocessed to determine the most significant factors in the GSD domain. To do this, the following three steps were followed:

1) *Merging step*: There were several duplicates; therefore, the first step in removing these duplicated factors was to merge the duplicates, which helped lower the total number by more than half.

2) *Filtration step*: After the merging stage, the factors were ranked and filtered by selecting only those with a frequency rate of greater than 50 percent. In this manner, the top 18 factors that affect software in the GSD domain can be collected.

3) *Categorization step*: In this phase, the top 18 factors were categorized into four categories: requirements, management, technical, and cultural, as shown in Table II to answer RQ1.

#### B. Dataset Selection

This subsection describes the data collection procedure and descriptive analysis of the dataset used to construct the model. The dataset was collected through a questionnaire survey and interviews with software companies and experts from various global regions. The main target was to focus on organizations that had extensive experience with outsourcing and were already using the GSD method. The dataset consists of information from 140 software projects in the GSD domain. Then, the data was gathered by conducting a questionnaire survey and interviews with companies and experts. The questionnaire is based on the top 18 factors listed in Table II. The data were collected from various regions that support a wide range of clients and vendors, including western Europe, central and eastern Europe, Africa, and the Middle East.

The dataset attributes were project ID, region, job, experience, company type, requirement clarity, project scope, requirement changes, project planning, project size, project management, communication, cost, commitment, modern technologies, roles and responsibilities, skilled staff, organizational stability, language and culture, time difference, progress, team size, methodology, and project status.

Then, the attributes were classified into numerical and categorical categories. The independent and dependent attributes were then determined. The dependent attribute is "Project Status." The remaining 23 attributes were independent attributes. Table III presents a more detailed description of the attributes of the dataset. In addition, Appendix "II" provides a sample of the questionnaire with attributes represented as questions.

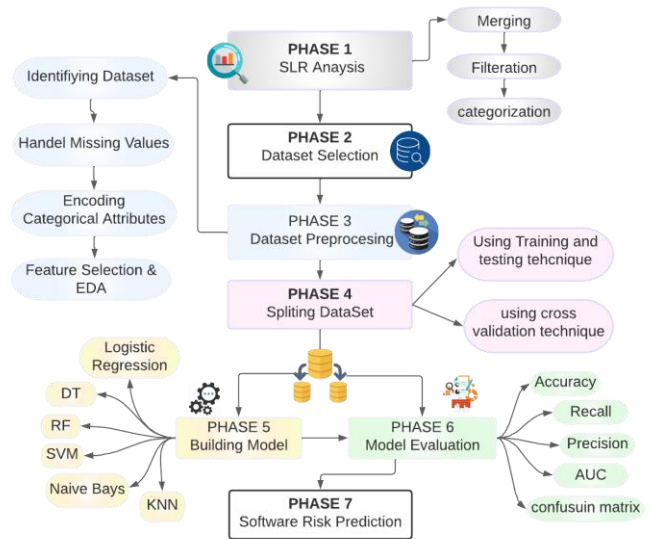


Fig. 2. A Proposed Model for Software Risk Prediction.

TABLE II. TOP 18 SOFTWARE FACTORS THAT AFFECT GSD

Requirement Factors	
1.	Requirement ambiguity
2.	Requirement changes
3.	Requirements scope
4.	New technologies
5.	Project size
Management Factors	
6.	Competence level of project manager
7.	No planning or inadequate planning
8.	Low commitment of stockholders
9.	Progress measure
10.	Cost balance
11.	Lack of roles and responsibilities
12.	Team size
Technical factors	
13.	Staff does not have required skills
14.	Unstable organizational environment
15.	Methodology followed
16.	Communication infrastructure and process
Cultural factors	
17.	Language and culture differences
18.	Time zone difference

TABLE III. DATASET VARIABLES AND DESCRIPTIVE STATISTICS

n	Attributes	Type	Description	Mean	Std	Min	50%	Max
1	Project ID	Numerical (int64)	Project unique ID, which starts with one and ends by 140	—	—	1	—	140
2	Region	Categorical (object)	Region attributes lie in 5 main regions: Western Europe, Central, and Eastern Europe, Asia, Africa, and the Middle East.	—	—	—	—	—
3	Job	Categorical (object)	The job role of the person/company developer who filled the form	—	—	—	—	—
4	Experience	Numerical (int64)	Team/individual experience measured in years	3.925	4.171	1	3	30
5	Company Type	Numerical (int64)	company types measured by (national, international, and startup)	0.535	0.528	0	1	2
6	Requirement Clarity	Numerical (int64)	the level of requirements clearness is measured as (clear, moderate, unclear, and ambiguous)	2.1	0.798	1	2	4
7	Project Scope	Numerical (int64)	the project scope is measured as (clear, moderate, unclear, and ambiguous)	2.071	0.810	1	2	4
8	Requirement Changes	Numerical (int64)	the project scope is measured as (minor, normal, heavy, and messy)	2.55	0.798	1	2	4
9	Project Planning	Numerical (int64)	the project planning is measured as (clear, moderate, unclear, and ambiguous)	2.214	0.863	1	2	4
10	Project Size	Numerical (int64)	the project size is measured as (Enterprise, large, medium, and small)	2.185	0.918	1	2	4
11	Project Management	Numerical (int64)	Project manager's quality is measured as (Expert, Moderate, Basic, and None)	2.142	0.844	1	2	4
12	Communication	Numerical (int64)	communication is measured as (Excellent, Moderate, need enhancements, and worst)	2.028	0.804	1	2	4
13	Cost	Numerical (int64)	cost is measured as (balanced, moderate, and not balanced)	1.842	0.626	1	2	3
14	Commitment	Numerical (int64)	stakeholders' commitment is measured as (High, moderate, and low)	1.75	0.669	1	2	3
15	Modern Technologies	Numerical (int64)	modern technologies are measured as (many, normal, and few)	1.75	0.613	1	2	3
16	Roles and Responsibilities	Numerical (int64)	responsibilities are measured as (clear, moderate, and unclear)	1.721	0.74	1	2	3
17	Skilled Staff	Numerical (int64)	skilled staff are measured as (Agree, moderate, and disagree)	1.485	0.64	1	1	3
18	Organization Stability	Numerical (int64)	organizational stability is measured as (stable, normal, and unstable)	1.607	0.716	1	1	3
19	Language and Culture	Numerical (int64)	language and culture are measured as (reasonable, can be handled, and unreasonable)	1.55	0.627	1	1	3
20	Time Difference	Numerical (int64)	Time Difference is measured as (reasonable, can be handled, and unreasonable)	1.7	0.675	1	2	3
21	Progress	Categorical (object)	Progress level Measured by (task level, module level, sprint level, and delivery level)	—	—	—	—	—
22	Team Size	Numerical (int64)	Team size measured by team members	0.807	0.855	0	1	4
23	Methodology	Categorical (object)	the methodology was measured as (waterfall, scrum, Kanban, extreme programming, feature-driven, lean development, crystal, and dynamic system development, and rapid development )	—	—	—	—	—
24	Project status	Numerical (int64)	Project status represents this project is success or failed	0.814	0.39	0	1	1

### C. Dataset Preprocessing

In this subsection, the data preprocessing techniques are presented. This phase can be considered as the initial phase for building the machine learning model. Real-world data are often incomplete, inconsistent, or incorrect (because they have outliers or mistakes). Thus, preprocessing techniques must be conducted to help refactor the dataset to keep it clean, formatted, and organized [24]. This subsection includes four steps of dataset preprocessing: identifying the dataset, finding,

and handling missing values, encoding categorical attributes, and feature selection.

1) *Identify dataset:* During data preparation, it is essential to identify insights into the dataset because improper handling may lead to misleading software model results and serious model risks. Table III shows that the dataset is divided into two main types: categorical and numerical. It also provides a full picture of the dataset's characteristics, such as its type,



description, mean, standard deviation (Std), and minimum and maximum values.

2) *Finding and handling missing attributes:* Incomplete data can lead to inaccurate results. Consequently, these situations may be addressed by finding the mean of the attributes using numerical data. This is more efficient than the usual methods of treating missing values, which include omitting the entire row or column, as this might lead to data misrepresentation or bias in the dataset. Alternatively, mean, median, or mode can be used.

3) *Encoding categorical data:* As it is known, machine learning deals with numerical attributes only. Thus, categorical attributes can't be used until they are transformed into numerical data. As a result, only four categorical attributes which are: "Region," "Job," "Progress," and "Methodology" should be transformed into numerical attributes. The Python scikit-learn library label encoder technique was used to transform the categorical attributes into numerical attributes. In this technique, each label is assigned a unique integer based on the alphabetical ordering [25].

4) *Feature selection:* From the list of 24 attributes in Table III, Independent attributes that are significant to the model must be chosen. Therefore, weak attributes or attributes that do not have a relationship with the model should be excluded. To determine these relationships, a correlation analysis was conducted, which is a common multivariate (EDA) that relies on statistical techniques to measure the linear relationship between attributes and each other to obtain a better insight into the factors and their relationships [26]. The correlation coefficient is the unit of measurement used to calculate the intensity between the two variables. It has three types:

- a) *Positive correlation:* (0 to 1) means that both attributes are in the same direction; an increase in one will increase the other, and vice versa
- b) *Negative correlation:* (-1 to 0) means that both attributes move in the opposite direction; an increase in one will decrease the other, and vice versa.
- c) *Weak/zero correlation:* (0) means that the two attributes do not affect each other.

The formula for the correlation coefficient can be written as:

$$R = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}} \quad (1)$$

where R is the correlation coefficient, usually from -1 to 1,  $x_i$  is the value from the X dataset,  $\bar{x}$  is the mean value of the X dataset,  $y_i$  is the value from the Y dataset, and  $\bar{y}$  is the mean value of the Y dataset. More details regarding the dataset attribute correlation are presented in Fig. 3.

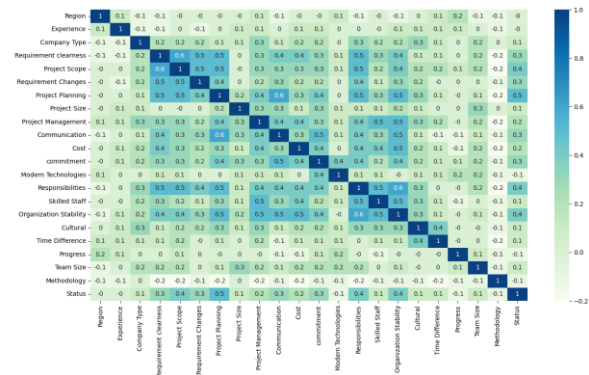


Fig. 3. Correlation Analysis for Dataset Attributes.

Fig. 4 represents the independent attributes that have  $R \geq 0.4$  ("Project Scope," "Project Planning," "Responsibilities," and "Skilled Staff") and independent attributes that have  $R \leq 0$  ("Region" and "Experience"), which will be excluded from the dataset so as not to affect the proposed model.

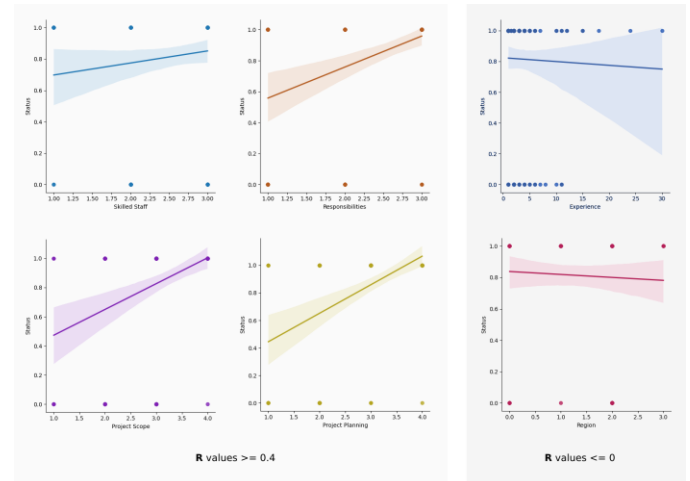


Fig. 4. Correlation Coefficient (R) between Attributes.

#### D. Machine Learning Model Builder

In this subsection, the implementation of ML models is discussed. The category for the proposed ML model is called a "supervised problem" because the dataset is labeled (one with the correct answer) which can be used to teach the model how to predict software risk [27]. The model is based on the top six machine-learning classification algorithms: logistic regression, (DT), (RF), support vector machine (SVM), K-nearest neighbor (KNN), and naïve Bayes.

1) *Logistic regression:* The logistic regression algorithm is a classification technique based on statistical procedures. Logistic regression is a widely used ML algorithm for binary classification that makes predictions based on the sigmoid function [28].

The sigmoid function can be defined as a mathematical procedural function that takes any real number and maps it to a probability between one and zero. The formula for the sigmoid function can be expressed as:

$$\sigma(x) = \frac{1}{1+e^{-x}} \quad (2)$$

where  $\sigma(x)$  is the sigmoid function that returns values ranging from zero to one,  $x$  represents the sample,  $e^{-x}$  represents the inverse of the exponential function  $\frac{1}{e^x}$

2) *Decision tree (DT)*: The decision tree algorithm can be represented as a hierarchical or flowchart which represents the data with decisions [25]. The decision tree has many branches created by splitting the dataset into subsets based on the essential attributes, and each branch can be considered as an if-else statement. To create the hierarchical structure in the decision tree, the Gini index algorithm was used to select the best attribute selection measures (ASM) to split the data. The Gini index algorithm can be written as:

$$Gini(D) = 1 - \sum_{i=1}^c P_i^2 \quad (3)$$

where  $c$  is the total number of classes, and  $P_i$  is the probability of picking the data point with class  $i$

3) *Random forest (RF)*: An RF is a collection of decision trees. It is a common ensemble method that aggregates the results of multiple models. RF uses the bagging technique, which allows each tree to be trained on random dataset sampling and takes the majority vote from the trees [25].

4) *Support Vector machine (SVM)*: SVM is a machine learning algorithm that can be used for classification and regression analysis [26]. The purpose of SVM is to classify data based on hyperplanes in an N-dimensional (number of attributes) space, which is the border between positive and negative classes, maximizing the distance between data points from different classes.

5) *K-Nearest Neighbour (KNN)*: The KNN is an analogy-based ML algorithm. In general, it uses the Euclidean distance to calculate the distance between points and each other and then assigns the label of new data based on the labels of the nearest data points. The Euclidean distance can be written as

$$d(y, x) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2} \quad (4)$$

where  $d$  is the Euclidean distance;  $(y, x)$  is the two-point Euclidean N-space;  $x_i, y_i$  represent the Euclidean vectors;  $n = N$ -space (attribute numbers).

6) *Naïve bays*: The naive Bayes algorithm depends on Bayes' theorem, which describes the probability of an event based on prior knowledge. Naive Bayes assumes that each feature is independent of the other [27]. The calculation of naive Bayes can be represented as:

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)} \quad (5)$$

where  $P(A|B)$  is a conditional probability, that is, the probability of an event A occurring given that B is true.

$P(B|A)$  is also a conditional probability: the probability of event B occurring given that A is true,  $P(A)$  and  $P(B)$  are the probabilities of observing A and B, respectively, without any conditions.

#### IV. EXPERIMENTAL EVALUATION

For model evaluation, different techniques and algorithms are described in this section. The essential part of any model is to determine whether it is accurate. Five evaluation metrics were used to measure the confusion matrix, accuracy, recall, precision, and area under the curve (AUC).

1) *Confusion matrix*: A confusion matrix is the best way to solve binary classification problems [35] because it shows the actual and predicted values and summarizes them in a matrix, as shown in Table IV.

2) *Accuracy*: Accuracy is the most important indicator for measuring a model's performance [29]. The purpose of the accuracy was to measure the percentage of the total number of correctly classified examples predicted over the total number of examples. The metric equation can be written as.

$$ACC = \frac{TP+TN}{TP+TN+FP+FN} \quad (6)$$

3) *Recall*: The recall evaluation metric, also known as the true positive rate (TPR), is used to determine the proportion of correctly classified positive classes [29]. The metric equation can be written as

$$R(TPR) = \frac{TP}{TP+FN} \quad (7)$$

4) *Precision*: The primary purpose of precision metrics is to measure the positive patterns from the total predicted patterns in a positive class [29]. The metric equation can be written as

$$P = \frac{TP}{TP+FP} \quad (8)$$

5) *AUC*: The area under the ROC curve (AUC) is a popular metric for comparing and optimizing machine-learning models [25]. A higher AUC indicates a better model performance. For classification evaluation, the AUC is more accurate than the accuracy metric, although the computational cost is high compared to the accuracy metric [25]. The AUC metric equation can be expressed as follows:

$$AUC = \frac{S_p - T_p(T_n + 1) / 2}{T_p T_n} \quad (9)$$

where  $S_p$  is the summation of all the positive examples,  $T_p$  is the number of positive examples, and  $T_n$  is the number of negative examples.

TABLE IV. CONFUSION MATRIX VALUE SUMMARIZATION

	Predicted (0)	Predicted (1)
Actual (0)	True Negative (TN)	False Positive (FP)
Actual (1)	False Negative (FN)	True Positive (TP)



V. RESULT AND DISCUSSION

In this section, the results of the software risk prediction model were discussed using six classification machine learning algorithms: logistic regression, DT, RF, SVM, (KNN), and naïve Bayes, and by using the dataset of 140 software projects in the real industry of global software development. Algorithm I present the pseudocode for the risk-prediction model using training data of 80% and 20% of the testing data.

**Algorithm I:** Pseudo-Code for Risk Prediction Model

- Input:** Import the dataset from a CSV File.
- 1: **Data Preprocessing Phase:** [ data cleaning, missing values]
  - 2: **Feature Transformation and Categorical Feature Encoding:**
  - 3: **Apply EDA and Feature Selection**
  - 4: **Dataset split: 80% for training and 20% for testing.**
  - 5: **Set:** Model = Logistic Regression, SVM, KNN, DT, RF, and Naïve Bayes
  - 7: **for each Model do**
  - 8:     **Select:** the ML model to use
  - 9:     **Use:** the training dataset to feed the proposed model
  - 10:    **Apply:** Testing the model using a training dataset
  - 11:    **Calculate:** confusion matrix
  - 12:    **Calculate:** The Accuracy metrics
  - 13:    **Calculate:** The Recall metrics
  - 14:    **Calculate:** The Precision metrics
  - 15:    **Calculate:** the AUC metrics
  - 16: **end for**

The software risk prediction model was constructed using the top six classification techniques: logistic regression, SVM, KNN, DT, RF, and naïve Bayes. Five evaluation metrics were used to find the most optimal ML algorithms to fit the risk prediction model. The model was conducted using the programming language Python and other third-party packages such as NumPy, Pandas, Scikit-Learn, Pandas, Matplotlib, and Seaborn, running on the MacBook Pro with the following specifications: Intel Core i5, 2.0Ghz, 16GB, and 512GB SSD. Table V presents a comparison of the six ML classification techniques using different evaluation metrics. Finally, the following research question was answered:

RQ2: What are the best machine learning techniques for software risk prediction in GSD?

Table V indicates that the top three techniques with the highest accuracy, AUC, recall, and precision were logistic regression, random forest, and SVM, with accuracy percentages of 89%, 85%, and 82%; AUC percentages of 73%, 71%, and 48%; recall percentages of 96%, 92%, and 96%; and precision percentages of 92%, 92%, and 85%, respectively.

TABLE V. SUMMARIZATION OF CONFUSION MATRIX VALUES

Model Name	Accuracy	AUC	RECALL	Precision
Logistic Regression	0.89	0.73	0.96	0.92
SVM	0.82	0.48	0.96	0.85
KNN	0.71	0.52	0.79	0.86
DT	0.71	0.62	0.75	0.90
Random Forest	0.85	0.71	0.92	0.92
Naïve Bayes	0.71	0.62	0.75	0.90

Therefore, logistic regression can be considered the optimum ML algorithm for software risk prediction in the domain of GSD, with an accuracy rate of approximately 90%. Further details regarding the algorithm’s confusion matrix showing the four values of true positives, true negatives, false positives, and false negatives are shown in Fig. 5.

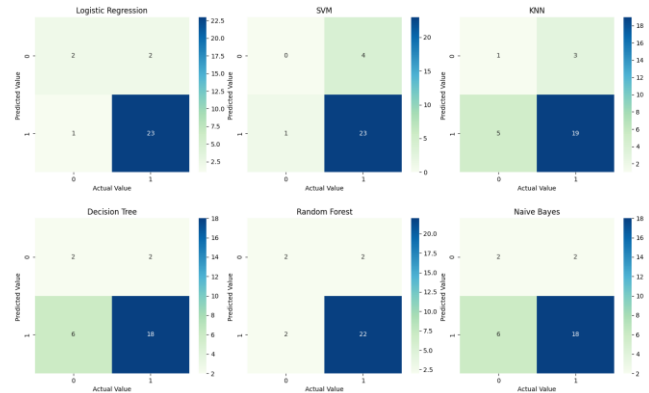


Fig. 5. The ML Confusion Matrix for the Six Algorithms.

In addition, to improve the results of other algorithms, another technique was applied for splitting the dataset called cross-validation, which is a statistical method for splitting data to test and train a model on different iterations. In other words, cross-validation split the training dataset into k smaller sets. This technique helped us improve the accuracy of most of the six algorithms, obtain better insights, and solve overfitting classification problems.

Table VI Shows a comparison of the six ML classification techniques after applying the cross-validation technique using five k-fold; four of them were used for the model training, and the remaining fold was used for validating the model.

TABLE VI. ML ALGORITHMS ARE SUBJECTED TO CROSS-VALIDATION

Model Name	Accuracy	AUC	RECALL	Precision
Logistic Regression	0.80	0.72	0.92	0.85
SVM	0.81	0.75	1	0.81
KNN	0.80	0.74	0.93	0.85
DT	0.78	0.65	0.85	0.90
Random Forest	0.80	0.81	0.92	0.84
Naïve Bayes	0.80	0.84	0.84	0.91

After applying the cross-validation technique, the accuracy of the KNN, DT, and naïve Bayes were increased by 9% to reach 80%. Thus, from Table VI can be observed that the accuracy of the six algorithms is approximately 80%, and the top three algorithms are Support Vector Machine, KNN, and logistic regression, with accuracy percentages of 81, 80, and 80%, respectively.

## VI. THREATS TO VALIDITY

This section discusses the reality of the study, based on internal and external threats and construct validity [30].

Internal validity relates to whether the investigated software risk prediction model is affected by other factors, such as Python Scikit-learn library parameters. Unfortunately, there is no standard method to choose this parameter, but the standard parameters and best practices was used in the Scikit-learn library to solve this problem [25]. The standard parameters, best practices, and configuration related to ML implementation for the six classification algorithms are provided in Appendix "III". Another internal threat is the split of the dataset; the dataset was divided into training and test sets at proportions of 80% and 20%, respectively. Random assignments were avoided to avoid influencing the model results. In addition, another preferred technique was used, called cross-validation, which splits the dataset into smaller datasets to train and test the model and calculate the average of these results to determine the most accurate result for the risk prediction model.

External validity is related to the generalization of the software risk-prediction model [30]. Dataset samples were tried to obtain from most of the companies that outsource and apply the GSD concept in different regions; however, with a limited number of datasets, some difficulties in generalizing the findings appear. In addition, there were difficulties in obtaining the data set because outsourcing companies often had to pay for their data, which limited the size of the dataset that could use.

The final threat is related to the reliability of the proposed model, this point was considered when conducting the model to validate and analyze its performance using the confusion matrix, accuracy indicator, recall metric, precision metrics, and AUC metrics. Also, these results were compared with those of the top six classification algorithms to determine the best one for the proposed model.

## VII. CONCLUSION

Obtaining a solid and accurate software risk-prediction model has always been difficult in global software development. The applied model will help software companies, experts, project managers, and developers predict software risk, which will reduce the amount of time and money spent on this approach.

A dataset of 140 software projects in different regions was used to build the model, and was collected using 18 software factors, which were carefully collected from past studies and reviewed by experts. The data preprocessing phase consisted of four steps: identifying the dataset, handling missing values,

encoding categorical attributes, feature selection, and conducting EDA analysis. Two techniques were used for the dataset splitting. The first technique is the common traditional technique, which uses 80% for training and 20% for testing without using any random or shuffle to avoid influencing the results of the model. The other technique is cross-validation using 5-k folds, with 4-folds used for model training and the remaining used to validate the model.

The results show that the top two algorithms were logistic regression and random forest with accuracy percentages of 89% and 85%, respectively. Also, cross-validation was used technique to improve the accuracy of the other models by approximately 80% and obtained better results.

## VIII. FUTURE WORK

Below are suggestions to improve the proposed model and the dataset that can be considered in the future:

- 1) The dataset needs to be expanded, by gathering the dataset from distributed locations that adopt the GSD approach.
- 2) Generalization of the findings.
- 3) Enhancing the accuracy of the ML algorithms.

## REFERENCES

- [1] J. Menezes, C. Gusmão, and H. Moura, "Risk factors in software development projects: a systematic literature review," *Software Quality Journal*, vol. 27, no. 3. Springer New York LLC, pp. 1149–1174, Sep. 01, 2019. doi: 10.1007/s11219-018-9427-5.
- [2] S. Ali, H. Li, S. U. Khan, M. F. Abrar, and Y. Zhao, "Practitioner's view of barriers to software outsourcing partnership formation: An empirical exploration," *Journal of Software: Evolution and Process*, vol. 32, no. 5, May 2020, doi: 10.1002/smr.2233.
- [3] A. Iftikhar, S. Musa, M. Alam, M. M. Su'ud, and S. M. Ali, "A Survey of Soft Computing Applications in Global Software Development," in 2018 IEEE International Conference on Innovative Research and Development (ICIRD), 2018, pp. 1–4.
- [4] University of Management and Technology (Pakistan), Institute of Electrical and Electronics Engineers. Lahore Section., and Institute of Electrical and Electronics Engineers, 3rd International Conference on Innovative Computing (ICIC): (IC)2 2019: 1st-2nd November 2019, Lahore, Pakistan.
- [5] J. A. Khan, S. U. R. Khan, J. Iqbal, and I. U. Rehman, "Empirical Investigation about the Factors Affecting the Cost Estimation in Global Software Development Context," *IEEE Access*, vol. 9, pp. 22274–22294, 2021, doi: 10.1109/ACCESS.2021.3055858.
- [6] M. A. Akbar, J. Sang, Nasrullah, A. A. Khan, M. Shafiq, and Fazal-E-Amin, "Towards the Guidelines for Requirements Change Management in Global Software Development: Client-Vendor Perspective," *IEEE Access*, vol. 7, pp. 76985–77007, 2019, doi: 10.1109/ACCESS.2019.2918552.
- [7] M. A. Akbar, M. Shafiq, T. Kamal, and M. Hamza, "Towards the successful requirements change management in the domain of offshore software development outsourcing: Preliminary results," *International Journal of Computing and Digital Systems*, vol. 8, no. 3, pp. 205–215, May 2019, doi: 10.12785/ijcds/080301.
- [8] M. Rizwan, J. Qureshi, A. Al-Zaidi, and R. Qureshi, "Global Software Development Geographical Distance Communication Challenges Article in International Arab Journal of Information Technology," 2017. [Online]. Available: <https://www.researchgate.net/publication/308952993>
- [9] A. Iftikhar, M. Alam, R. Ahmed, S. Musa, and M. M. Su'ud, "Risk Prediction by Using Artificial Neural Network in Global Software Development," *Comput Intell Neurosci*, vol. 2021, pp. 1–25, Dec. 2021, doi: 10.1155/2021/2922728.

- [10] M. Yaseen and Z. Ali, "Success Factors during Requirements Implementation in Global Software Development: A Systematic Literature Review," *International Journal of Computer Science and Software Engineering (IJCSSE)*, vol. 8, no. 3, 2019, [Online]. Available: [www.IJCSSE.org](http://www.IJCSSE.org)
- [11] B. J. Galli, "Addressing Risks in Global Software Development and Outsourcing," *Int J Risk Conting Manag*, vol. 7, no. 3, pp. 1–41, May 2018, doi: 10.4018/ijrcm.2018070101.
- [12] Asim Iftikhar, Muhammad Alam, Shahrulniza Musa, and Mazliham Mohd Su'ud, "Trust Development in Virtual teams to Implement Global Software Development (GSD): A Structured Approach to Overcome Communication Barriers," in *2017 IEEE 3rd International Conference on Engineering Technologies and Social Sciences (ICETSS)*, 2017, pp. 1–5.
- [13] B. G. Tavares, C. E. S. da Silva, and A. D. de Souza, "Risk management analysis in Scrum software projects," *International Transactions in Operational Research*, vol. 26, no. 5, pp. 1884–1905, Sep. 2019, doi: 10.1111/itor.12401.
- [14] Project Management Institute, *A guide to the Project Management Body of Knowledge (PMBOK guide)*, 6th ed. Newton Square, PA: Project Management Institute, 2017.
- [15] A. S. Filippetto, R. Lima, and J. L. V. Barbosa, "A risk prediction model for software project management based on similarity analysis of context histories," *Inf Softw Technol*, vol. 131, Mar. 2021, doi: 10.1016/j.infsof.2020.106497.
- [16] M. A. Akbar, M. Shameem, A. A. Khan, M. Nadeem, A. Alsanad, and A. Gumaei, "A fuzzy analytical hierarchy process to prioritize the success factors of requirement change management in global software development," *Journal of Software: Evolution and Process*, vol. 33, no. 2, Feb. 2021, doi: 10.1002/smr.2292.
- [17] M. Marinho, R. Camara, and S. Sampaio, "Toward unveiling how safe framework supports agile in global software development," *IEEE Access*, vol. 9, pp. 109671–109692, 2021, doi: 10.1109/ACCESS.2021.3101963.
- [18] S. K. Niranjana, V. N. M. Aradhya, Amity University, IEEE-USA, Institute of Electrical and Electronics Engineers. Uttar Pradesh Section, and Institute of Electrical and Electronics Engineers, *Proceedings of the 2016 2nd International Conference on Contemporary Computing and Informatics (IC3I)*: 14-17 December 2016, Noida, India.
- [19] A. N. Okon, S. E. Adewole, and E. M. Uguma, "Artificial neural network model for reservoir petrophysical properties: porosity, permeability and water saturation prediction," *Model Earth Syst Environ*, vol. 7, no. 4, pp. 2373–2390, Nov. 2021, doi: 10.1007/s40808-020-01012-4.
- [20] K. Sahu and R. K. Srivastava, "Soft computing approach for prediction of software reliability," *ICIC Express Letters*, vol. 12, no. 12, pp. 1213–1222, Dec. 2018, doi: 10.24507/icicel.12.12.1213.
- [21] K. Suresh and R. Dillibabu, "An integrated approach using IF-TOPSIS, fuzzy DEMATEL, and enhanced CSA optimized ANFIS for software risk prediction," *Knowl Inf Syst*, vol. 63, no. 7, pp. 1909–1934, Jul. 2021, doi: 10.1007/s10115-021-01573-5.
- [22] J. A. Khan, S. U. R. Khan, T. A. Khan, and I. U. R. Khan, "An Amplified COCOMO-II Based Cost Estimation Model in Global Software Development Context," *IEEE Access*, vol. 9, pp. 88602–88620, 2021, doi: 10.1109/ACCESS.2021.3089870.
- [23] M. Assim, Q. Obeidat, and M. Hammad, "Software Defects Prediction using Machine Learning Algorithms," Oct. 2020. doi: 10.1109/ICDABI51230.2020.9325677.
- [24] Jacqueline Kazil and Katharine Jarmul, *Data Wrangling with Python: Tips and Tools to Make Your Life Easier*. O'Reilly Media, Inc. 2016.
- [25] E. Bisong, *Building Machine Learning and Deep Learning Models on Google Cloud Platform*. Apress, 2019. doi: 10.1007/978-1-4842-4470-8.
- [26] H. D. P. de Carvalho, R. Fagundes, and W. Santos, "Extreme Learning Machine Applied to Software Development Effort Estimation," *IEEE Access*, vol. 9, pp. 92676–92687, 2021, doi: 10.1109/ACCESS.2021.3091313.
- [27] R. Saravanan and Pothula Sujatha, "A State of Art Techniques on Machine Learning Algorithms: A Perspective of Supervised Learning Approaches in Data Classification," 2018.
- [28] Hilbe and J.M., *Practical Guide to Logistic Regression*; Chapman and Hall/CRC: Boca Raton, FL, USA. 2016.
- [29] H. M and S. M.N, "A Review on Evaluation Metrics for Data Classification Evaluations," *International Journal of Data Mining & Knowledge Management Process*, vol. 5, no. 2, pp. 01–11, Mar. 2015, doi: 10.5121/ijdkp.2015.5201.
- [30] D. A. Broniatowski and C. Tucker, "Assessing causal claims about complex engineered systems with quantitative data: internal, external, and construct validity," *Systems Engineering*, vol. 20, no. 6, pp. 483–496, Nov. 2017, doi: 10.1002/sys.21414.

#### APPENDIX

1) factors that affect software in the GSD and explaining the three steps [collected, merged, and filtered] applied to these factors: [https://github.com/AI3ameed/ML\\_Classification\\_GSD/blob/main/software%20factors.docx](https://github.com/AI3ameed/ML_Classification_GSD/blob/main/software%20factors.docx)

2) An Example of the Questionnaire form (published on GitHub): [https://github.com/AI3ameed/ML\\_Classification\\_GSD/blob/main/questionnaire\\_samples.zip](https://github.com/AI3ameed/ML_Classification_GSD/blob/main/questionnaire_samples.zip)

3) ML classification and a sample of the dataset that used for model prediction: [https://github.com/AI3ameed/ML\\_Classification\\_GSD](https://github.com/AI3ameed/ML_Classification_GSD)

# HelaNER 2.0: A Novel Deep Neural Model for Named Entity Boundary Detection

Y.H.P.P Priyadarshana, L Ranathunga

Department of Information Technology, University of Moratuwa, Moratuwa, Sri Lanka

**Abstract**—Named entity recognition (NER) is a sequential labelling task in categorizing textual nuggets into specific types. Named entity boundary detection can be recognized as a prominent research area under the NER domain which has been heavily adapted for information extraction, event extraction, information retrieval, sentiment analysis etc. Named entities (NE) can be identified as per flat NEs and nested NEs in nature and limited research attempts have been made for nested NE boundary detection. NER in low resource settings has been identified as a current trend. This research work has been scoped down to unveil the uniqueness in NE boundary detection based on Sinhala related contents which have been extracted from social media. The prime objective of this research attempt is to enhance the approach of named entity boundary detection. Considering the low resource settings, as the initial step, the linguistic patterns, complexity matrices and structures of the extracted social media statements have been analyzed further. A dedicated corpus of more than 100,000 tuples of Sinhala related social media content has been annotated by an expert panel. As per the scientific novelties, NE head word detection loss function, which was introduced in HelaNER 1.0, has been further improved and the NE boundary detection has been further enhanced through tuning up the stack pointer networks. Additionally, NE linking has been improved as a by-product of the previously mentioned enhancements. Various experimentations have been conducted, evaluated and the outcome has revealed that our enhancements have achieved the state-of-art performance over the existing baselines.

**Keywords**—Computational linguistics; deep neural networks; natural language processing; named entity boundary detection; named entity recognition

## I. INTRODUCTION

Named entity recognition (NER) is an active research area and a prominent computational task under the key area of natural language processing (NLP) [1]. Further, NER can be recognized as a semantic level sequential labelling task in terms of identifying textual spans into the specified categories such as Person, Location and Organization. Several textual spans can be categorized under the above-mentioned predefined categories which are known as named entities (NE). Named entities can be categorized into two main different aspects such as named entity type and the named entity boundary. When it comes for the NE type aspect, two variations can be established such as fine-grained NERs and coarse-grained NERs. Fine-grained NERs typically consider much larger scopes and its contribution towards the NE boundary aspect would be slightly low. When it comes for coarse-grained NERs, which is considered as high potential capabilities in demonstrating the NE boundary related aspects

on the basis of specific niche NE types. The specific niche NE types have already been taken into the consideration under the flat NE boundary aspects [2]. Considering the NE boundary and type aspects still there is a potential vacuum under NE boundary avenues compared to the other variation, NE type component.

A novel approach has been invented through a recent prominent research activity to detect the respective boundaries of named entities [2]. The solution has been focused on the domain of Sinhala statements which have been extracted from social media such as Facebook, YouTube, and Twitter. This solution has been further enhanced from HelaNER 2.0 where an enhanced algorithm has been demonstrated to determine the respective named entity mention boundaries. These novel enhancements would be worthwhile to showcase the importance of addressing the social linguistic issues specially for low resource language settings. Named entity boundary detection still can be recognized as an unknown area considering Sinhala language. In terms of manipulating the social media contents which have been published in Sinhala, it is really worthwhile to undercover these types of niche domains. Compared to the other available benchmarks for NE boundary detection, several novel avenues and enhancements have been identified and showcased from this research attempt. The main contributions of this paper consist of the following:

1) The size of the combinational multi-layer classifier which is dedicated for NE head word detection along with the NE type detection has been increased. Additionally, the NE head word detection loss function has been further improved. Due to these enhancements in HelaNER 2.0, the performance of detecting the NE head words has been increased.

2) The stack pointer network which has been used for NE boundary detection under [2], has been further tuned up and improved. Specially, the 2nd and 3rd hidden layers which are dedicated for deriving character level representations have been further enhanced. These enhancements have been critically demonstrated and evaluated in the respective sections of this paper. Due to this enhancement, the overall accuracy for NE boundary detection process has been increased in a considerable margin compared with the other existing benchmarks.

3) As a by-product, the overall NE linking process has been further improved due to the above-mentioned enhancements. These novel enhancements have been critically analyzed and showcased under the methodological section.

4) Those main contributions have been critically evaluated considering the state-of-the-art benchmarks. The corpus has been increased compared to our previous research attempt [2]. The results have been demonstrated under the experimental results section.

The ultimate motivation of this research attempt is to develop a novel framework for named entity boundary detection for social media analysis. Detecting both NE boundary and NE type for named entities as an aggregate mechanism will eventually tune up the accuracy and performance of NE linking to knowledge bases. This proposed novel framework allows to capture the named entity boundary detection along with their respective named entity types by filling out the research gap which has been identified where the overall solution is capable of handling multiple domains including Sinhala. This framework can be further introduced as a specialized niche version for the domain Sinhala and is capable of enhancing up to a more generalized version considering multiple domains.

The rest of this paper is organized as follows. Related work about named entity (NE) boundary detection along with the specific NE type is given in Section II. Section III presents the scientific methodology of the overall approach in detail manner. Section IV introduces the experimental results of this entire solution along with the comprehensive evaluation procedure which has been conducted considering the existing prominent baselines. The discussion and future avenues of this novel mechanism have been discussed in Section V.

## II. RELATED WORK

Multiple deep neural and non-neural based computational systems have been established to showcase the usage of detecting NER in social media analysis in the recent past. Some of such systems have shown promising results over the existing benchmarks. The overall literature review would be separated into two sections such as analyzing the existing NE boundary detection systems along with the respective algorithms and critically analyze the statements which have been circulated in the social media platforms.

### A. Analyzing NE Boundary Detection

Due to the low language resources settings, deep transfer learning has been adapted to determine the starting and ending indices of name entities [3]. Such a system has been introduced to overcome certain issues such as removing the noisy data and avoiding hand-crafted feature settings. Due to the noisiness of the corpus, the system has failed in determining the respective NE types along with detecting the boundaries. In terms of enhancing NER, multiple NE linking mechanisms of detecting boundaries have been introduced considering the lexical and morphosyntactic features [4]. Classifying textual segmentations into text span identification has been identified as a significant drawback of this approach. A novel stack pointer networks-based approach along with deep adversarial transfer learning which is capable of detecting the NE boundaries has been demonstrated [5]. This system has a unique capability of deriving both starting and ending NE boundary tags but still has failed in detecting the respective NE type. In terms of addressing the issue of [6], a

context encoding neural model has been invented [7]. A combination of Bi-LSTM (Long Short-Term Memory) model with a CNN (Convolutional Neural Network) to capture character-level features relates to NE type and boundary aspects can be identified as the main contribution of the above-mentioned approach. Even though this model has performed well in capturing NE boundaries still the NE type capturing has to be improved further.

BDRYBOT, a neural based model for detecting NE boundaries has been demonstrated considering the NE linking procedures [1]. The model has been consisted with a CNN in terms of enhancing the character level contextual representations. The encoding phase has been enriched with a Bi-GRU (Gated Recurrent Units) model instead of Bi-LSTM considering low computational consumption. When it comes for the NE type detection along with NE boundary detection, still the system has not been improved to achieve up to that level and can be stated as a future avenue. A novel multitasking learning neural based boundary detection model has been introduced considering the nature of inner and outer layers of nested entity mentions [8]. This approach has been enriched with sequential classification tasks in terms of reducing the computational costs hence it has demonstrated some promising results in nested NER over the existing benchmarks. Even though this model has shown some promising results, still it has been limited for extracting implicit NE mention regions. Another hybrid NE tagging architecture has been showcased considering dual neural models, Bi-LSTM, and stack LSTM [9]. Here, a special NE tagging pattern, IOBES scheme has been adapted [10]. Though the NE identification performance has been boosted up due to this novel NE tagging pattern still this has been limited to NE explicit variations.

A novel deep neural network based exhaustive approach has been invented for nested NE mention recognition [11]. Here, both NE boundary related information and NE type related information have been considered. Additionally, inner NE feature representation also has been considered under this approach. Bi-LSTM model has been developed to fulfill the objectives by visualizing the character-based feature representations on top of contextual word vectors. Compared to the other well-known approaches which have been dedicated for nested NE tasks [12] [13] [14], the exhaustive model has shown promising outcomes under the time related complexities as well. Still there is a vacuum for entity mention outside feature representations apart from only limiting to inside feature mappings. A recent deep neural based approach, anchor region networks, has been demonstrated to showcase the value and impact of nested NEs [15]. The primary assumption has been set off as the head or main words would govern the whole structure of entity mention nuggets considering the respective boundary aspects. These fundamentals have been used to determine the benchmarking models with some major enhancements. Span based models [16] [17] [18] have been considered as a revolutionary approach for determining NER. Classification of sentence nuggets into the respective subsequences using span-based architecture [19] has been exhibited recently. Both multitask learning and BERT based classifications have been used to

play with the boundary level aspects of NEs. Even though this has been accepted as a novel methodology, still the conducted experiments have been revealed that the performance of detecting NE boundaries through span-based models is poor.

An enhanced Machine Reading Comprehension (MRC) model has been introduced in terms of addressing both flat NERs and nested NERs scenarios [20]. This novel model has focused on the theoretical approaches under the sequential labelling domain. In nature, MRC models are dedicated for extracting NE mention spans regardless of flat or nested nature [21]. The text classification objective has been accomplished using BERT based models [22]. Even though this model has been performed well enough for coarse-grained NER types, still there is a demand for such an advancement for considering fine-grained NER types as well. There is a trend of adapting object detection and mapping techniques in the theory of computer vision for extracting NE mentions. One of such advancements has been exhibited as NE boundary regression model [23]. In this approach, Nested NE mention detection has been conducted considering the overlapping of each NE mentions in the context by applying a concept called NE bounding boxes. A deep neural network model called convolutional feature maps has been invented for NE boundary detection. The same technique has been used for textual information extraction and information retrieval from movie reviews recently [24] [25]. Even though this model has showcased some promising results over the existing benchmarks still NE type detection aspect has been missed.

#### B. Analyzing Social Media Statements

Even though Sinhala is a morphological rich language, very limited amount of NLP resources has been introduced considering both micro level and macro level tasks in computational linguistics. Hence, very minimum facilities can be observed for accomplishing major activities like named entity boundary detection for Sinhala context. The paradigms of Sinhala social media statements analysis can be listed down as lexicon-based approaches and machine learning approaches [26]. The lexicon-based category has been specialized in applying for scenarios such as catering online forums, online blog posts and comments which have been extracted from online channels [27]. A specific speech lexicon has been adapted in terms of capturing verbs and nouns in the context [28]. Here, a NE recognizer has been used to detect the related nouns in the context. Since the whole procedure has been dominated under a particular lexicon, the approach can be recognized as a restrictive limited approach. In the recent past, an overwhelming demand has been experienced for the adaption of machine learning based models in capturing hate speech related contents. A novel approach has been introduced to pick out hate speech related content from German corpus [29]. This approach has been enriched with the adaption of transfer learning, web scrapping mechanisms and usage of Bag-Of-Words (BOW). Another similar approach has been conducted to extract speech related content from Indonesian language [30]. Here, a random forest classifier has been adapted to showcase the value of word n-gram feature representations in classifying social media statements.

When it comes for the avenues in deep learning for classifying social media statements, the linguistic level

matrices under micro level and macro level should be examined. A significant corpus is essential for implementing a valid supervised type of deep neural based model for accomplishing the analysis of extracted statements. Crowd sourcing can be identified as a handy approach to fulfill the issue of corpus unavailability [31] [32]. Those extracted data would be used to determine the respective word embeddings since it has been identified as the state of the art which has been discovered under textual processing [32] [33]. A unique classifier which has been invented adapting ensemble fundamentals has shown some promising results in classifying sexist and racist corpora [34]. Another novel LSTM based model has been introduced in terms of classifying speech related content in Italian language [35]. Those attempts can be identified as the fundamentals in deep neural models which have been invented for classifying social media content.

### III. RESEARCH METHODOLOGY

The overall methodological approach can be categorized into two main components such as discovering the Sinhala related posts, comments, and statements which have been spread out in social media context through mandatory features, complexity measurements and other related aspects and constructing a novel framework for capturing named entity boundary detection considering the respective named entity type.

#### A. Sinhala NE Boundary Detection

The ultimate goal is to turn up with a neural based methodology to discover and analyze the Sinhala related social media context analysis based on a supervised approach. As the initial phase, Sinhala social media related corpus has been constructed. A specific web crawling mechanism has been used to extract the social media statements from Facebook, Twitter, and YouTube platforms. Once the statements have been crawled, those have been collected to a pool to be distributed among multiple set of annotators who are responsible for conducting manual annotations. More than 100,000 social media statements have been crawled for this purpose. Once the annotated statements have been constructed, then the next sets of preprocessing tasks have been determined. Several important intermediate steps have been followed such as preprocessing, stemming, parts-of-speech (POS) tagging to filter the dataset in terms of using it in the next phases. Considering the Sinhala POS tagging, due to the unavailability of a standard Sinhala POS tag set, a special pseudocode has been implemented and demonstrated under the Fig. 1.

Then as per the final step under this section, the specific neural model has been designed and implemented. A recurrent neural network (RNN) LSTM has been used as the foundation for implementing the deep neural model since RNN has been accepted as the state-of-the-art in processing sequential representations. Some optimal hyper-parameter settings have been enriched for obtaining better results under the training procedure. The hyper-parameter settings can be identified as a vital component of the entire procedure where some of the critical evaluations have been conducted considering different value adjustments. The deep neural model which is dedicated for NE boundary detection can be showcased as per Fig. 2.



```

Algorithm 4.1: POS Tagging
SET chunker TO RegexpParser(r'''
NP:
{<NNPI>*<JJ>*<NNN>*<NNN><NNPA.*>*<NNPI.*>*<PRP.*>*}
VP:
{<JVB>*<NVB>*<V.*>*}
''')
FOR subtree IN parsed_tree.subtrees(filter=lambda t: t.label() EQUALS 'NP'):
    SET first_so TO []
    SET result TO subtree.leaves()
    OUTPUT(result)
    FOR word, tag IN result:
        first_so.append(word)
        SET so TO ''.join(first_so)
        so_list.append(so)
    OUTPUT('SO List: ' + str(so_list))
IF len(so_list) != 2:
    OUTPUT('System has not taken the Subject-Object correctly')
ELSE:
    SET sub TO so_list[0]
    SET ob TO so_list[1]
RETURN triple
    
```

Fig. 1. Pseudocode for Parts of Speech (POS) Tagging.

```

Algorithm 4.1: NE Identification
FOR max_len IN max_len arr:
    SET startTime TO time.time()
    SET tok TO Tokenizer(num_words=max_words)
    tok.fit_on_texts(X_train)
    SET train_sequences TO tok.texts_to_sequences(X_train)
    SET train_sequences_matrix TO sequence.pad_sequences(train_sequences,
maxlen=max_len)
    DEFINE FUNCTION RNN():
        SET INPUTs TO Input(name='INPUTs', shape=[max_len])
        SET layer TO Embedding(max_words, 50,
INPUT_length=max_len)(INPUTs)
        SET layer TO LSTM(10)(layer)
        SET layer TO Dense(256, name='FC1')(layer)
        SET layer TO Activation('relu')(layer)
        SET layer TO Dropout(0.5)(layer)
        SET layer TO Dense(1, name='out_layer')(layer)
    
```

Fig. 2. Pseudocode for LSTM based Deep Neural based Model for Sinhala NE Boundary Detection.

**B. Determining NE Boundary Detection**

Once the initial methodological step is ready by identifying and classifying the named entity identification for Sinhala, then the second phase which is determining the NE boundary detection can be initiated. The proposed architecture of the entire procedure for deriving named entity boundary detection can be showcased as Fig. 3. As per the word representation, word level embedding, and character level embedding have been followed. Since the embedding process is at a basic stage for Sinhala, Glove word embedding which has been developed based on English corpus will be used as the platform. The obtained embedding values have been fed to a Bi-LSTM layer in terms of generating the backward and forward value sets.

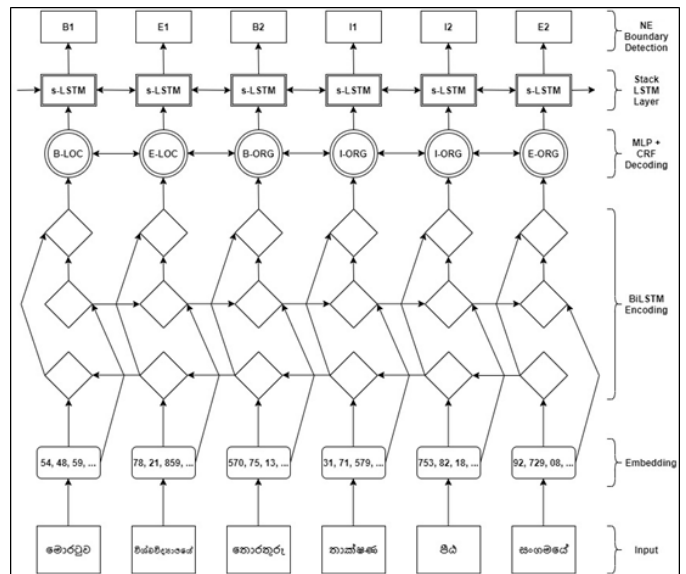


Fig. 3. Proposed Architecture for Named Entity Boundary Detection.

Special LSTM based deep neural model has been developed as the overall NE detection end to end framework. A Bi-LSTM layer would be used to derive the word level classifiers to obtain the respective head word which acts as the dominant words of the pre-defined context. As an example, considering English context, “Dilan Perera has been elected as the Secretary of the Colombo Sports Club”, Dilan Perera acts as a PER mention type while Colombo Sports Club takes ORG mention type. Additionally, Colombo could be recognized a LOC type. Considering the whole nugget, even though there are three main NE mentions, PER mention can be considered as the main or the head NE which governs the overall semantic contextual representation. So, the first step is to determine the head word of each unique sentence nugget. Once the head words have been obtained, the relevant contextual level representations have been fed into a structured classifier called multi-layer perceptron (MLP) along with the conditional random fields (CRF). MLP is capable in nature to derive the inner representations of the word level contextual representations. A combined classifier has been used to improve the level of accuracy and this combined mechanism can be mentioned as a unique approach which has not been used for any of the previous NE boundary detection frameworks. The whole process can be visualized as Fig. 4 as follows.

Once the head words have been obtained, the next step is to locate the specific entity mention nuggets per each dominant word. Here a novel theory called boundary bubbles has been introduced. Boundary bubbles (BB) are abstractive level contextual representations which are used to visualize the NE mention nuggets along with their respective identified NE head words. Fig. 5 shows the demonstration of boundary bubbles for the previously given example (Dilan Perera has been elected as the Secretary of the Colombo Sports Club).



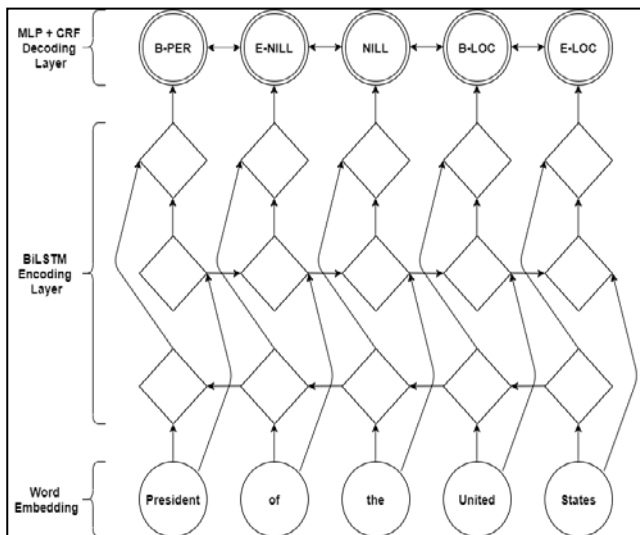


Fig. 4. Main System Architecture of NE Type Detection.

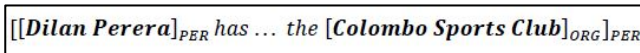


Fig. 5. Abstract Level Representation of Boundary Bubbles.

In terms of capturing mention nuggets, a special algorithm has been developed. Several parameters and assumptions have to be considered when declaring the specific algorithm. Two main such assumptions would be like, at least one word must be declared as the head word for that specific mention nugget and such a word should be found through the specific boundary bubble. The left boundary region and the right boundary region of the respective bubble should be determined for identifying the mention nuggets. Considering left and right boundary related values, two value tuples have been defined. A unique loss function has to be determined since it's mandatory to define the loss value under head word detection procedure. The head word detection loss function can be obtained as:  $-\log P(c_i|x_i)$ . The collective loss function would be determined as a weighted average loss value considering the respective scenarios including the head word detector would be resulting to NILL. The NE mention nuggets identifier can be recognized as per (1) as follows.

$$\tau(x_i; \emptyset) = v_i \cdot [-\log P(c_i|x_i) + L^R(x_i; \emptyset)] + (1 - v_i) \cdot [-\log P(NIL|x_i)] \quad (1)$$

$$L^R(x_i; \emptyset) = L^{left}(x_i; \emptyset) + L^{right}(x_i; \emptyset) \quad (2)$$

The respective above stated (2) would determine the identical structure for NE mention nuggets along with the respective starting and ending boundary regions. Additionally,  $v_i$  under (1), states the respective correlational value of the underline boundary bubble, in which the higher the value determines the stronger association considering the type of the bubble. The whole process would be complex if multiple head words per mention nuggets have been established. With all of these assumptions and decisions,  $v_i$  can be further demonstrated as:

$$v_i = \left[ \frac{P(c_i|x_i)}{\max_{x_t \in B_i} P(c_i|x_t)} \right] \alpha^v \quad (3)$$

Here  $\alpha$  can be showcased as a specific hyper-parameter. Different values can be assigned with the intention of analyzing the outcome behavior of the overall procedure. As the basic value assignment,  $\alpha = 0$  could generate a scenario where the determined NE mentions would be annotated with the respective boundary bubble type. When it comes for extracting the inner most head words,  $B_i$  determines the region of the extracted mention nuggets.

The final segment of the whole process of named entity boundary detection and type determination will be the boundary region classification. From the previous step, the respective NE mention nuggets have been derived and the next step is to determine the boundary detection using the obtained mention nuggets. In terms of deriving the NE boundary detection, a novel NE scope deriving mechanism which is called as IOBE (Inside, Outside, Beginning, End) pattern has been used. Even though IOB (Inside, Outside, Beginning) pattern has been adapted for NE boundary detection, usage of IOBE pattern can be identified as the first-time approach of adapting IOBE pattern along with stack pointer networks in terms of deriving the boundary detection. Plenty of other different patterns have been used for accomplishing different avenues under the named entity recognition and named entity boundary detection domains. Specially, BEO (Beginning, End, Outside) pattern has been used for one of the previous approaches under NE boundary assembling mechanisms for named entity recognition in biomedical domain [36]. As per our understanding, this is the initial attempt of using IOBE pattern for named entity boundary detection purposes under the named entity recognition domain. Further, this can be mentioned as one of the scientific core novelties of this research work. The respective design architecture related to NE boundary detection can be exhibited as follows under Fig. 6.

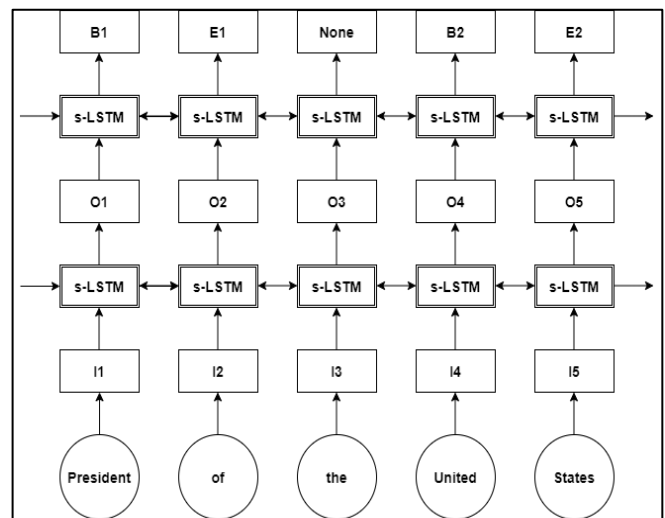


Fig. 6. Stack LSTM Architecture of NE Boundary Detection.

NE linking is a crucial task under information retrieval from the respective knowledge bases. Enhancing NE linking can be identified as a sub product of this overall procedure. Considering the three different components under the NE linking, our intention is to enhance the direct mapping through

the identification of NE boundaries. An explicit novelty will not be achieved from NE linking but a performance increment would only be expected due to the enhancements which have been exhibited under NE boundary detection. The respective enhanced algorithm which is dedicated for NE direct mapping purposes can be exhibited as (4). In the respective algorithm,  $p(c_{ij}|e_i)$  determines the respective contextual level probabilistic distribution of each entity mention denoted as  $e_i$ . The Laplacian smoothing has been adapted to avoid the zero probability values.

$$p(c_{ij}|e_i) = \frac{t(e_i, c_{ij}) + 1}{\#e_i + |C_i|} \quad (4)$$

The overall implementation has been carried out in a well constructive manner. Python 3.7.4 has been used as the primary programming language to implement the overall system including all the components. Frameworks such as TensorFlow 1.15 and Flask 2.1 have been adapted to deliver the whole application in a much standard way. The mandatory natural language processing and computational linguistics libraries such as Torch 1.7, Spacy 2.1.9, Keras 2.2.4, NLTK 3.4.5 have been adapted to fulfill the overall implementation procedure in a smooth manner. After the system has been implemented, the experimentation can be carried out, and the related experimental results can be described in detail manner as follows.

#### IV. EXPERIMENTAL RESULTS

The overall testing procedure has been conducted considering all the required aspects. An extensive experimental process has been carried out considering all the implemented components which have been mentioned earlier. Considering the overall experimental settings, several existing baseline mechanisms have been considered such as regional based models, conventional CRF models and hypergraph based computational models. A unique Sinhala NER corpus has been obtained which consists with 120,000 tuples as per training purposes, 100,000 tuples as per validation purposes and another 80,000 set as per accomplishing testing purposes. As per the hardware requirements, a high-performance graphics processing unit (GPU) machine has been acquired. Major sets of hyper-parameters have been set off for obtaining the competitive edge. Experimentations and the respective evaluation process have been focused with several sub-components such as evaluation on Sinhala NE identification, evaluation on word embedding, evaluation on NE head word detection, evaluation on NE mention boundary detection and evaluation on NE linking.

##### A. Evaluation on Sinhala NE Identification

Several annotators have been employed for constructing the main corpora such as classifying social media statements and grouping the identified named entities into NE categories. Hence, the Fleiss' Kappa [37] values have to be measured to evaluate the reliability of the corpus which has been used in the entire system. Once the annotation process has been accomplished, the inter annotator agreement has been processed to evaluate the relevant Kappa statistics. These respective individual Kappa statistics can be showcased under the Table I as follows.

TABLE I. KAPPA VALUES FOR NE CATEGORIES

Kappa Values for Individual NE Categories					
NE Class	Conditional Probability	Kappa	Standard Error	Z Value	P Value
PER	0.84	0.76	0.12	5.12	0.04
LOC	0.76	0.64	0.12	4.87	0.04
ORG	0.67	0.62	0.12	4.28	0.03
DES	0.48	0.57	0.12	3.72	0.02
PRO	0.43	0.52	0.12	3.24	0.02

Once the Kappa values have been obtained for the respective individual NE categories, then the overall Kappa statistics can be obtained and demonstrated as follows.

According to the values of Table II, the overall Kappa value for NE identification and categorization can be recognized as 0.72. Considering the overall value and according to the classification of the Fleiss' Kappa, it can be concluded that the generated Kappa value has represented a good strength in the inter annotator agreements.

TABLE II. OVERALL KAPPA VALUES FOR NE CATEGORIES

Overall Kappa					
NE Class	Conditional Probability	Kappa	Standard Error	Z Value	P Value
Overall	0.78	0.72	0.11	4.86	0.02

In terms of evaluating the model on NE type classification, set of main parameters have been defined and discussed in the previous chapter. The obtained results have been critically evaluated with the available benchmarks as per the Table III. The newly introduced unique model has been demonstrated as BB2022 (boundary bubbles).

TABLE III. EVALUATION ON NE TYPE CLASSIFICATION

Model	Precision (%)	Recall (%)	F1 (%)
SH2016 <sup>1</sup>	75.9	70.1	72.8
KBP2018	72.6	73.0	72.8
ARN2019	75.2	72.5	73.9
<b>BB2022</b>	<b>76.2</b>	<b>73.6</b>	<b>74.9</b>

As per the exhibited comparison outcomes, it can be concluded that the proposed novel model performs better than the existing NE type detection benchmarks in a competitive edge.

##### B. Evaluation on Word Embedding

Due to the low level of language resources settings, a transfer learning mechanism must be adapted to derive the Sinhala related word embeddings. Specific four main hyper-parameters have been used in terms of deriving the training procedure such as dev detect, test detect, dev all and test all. Once the overall testing process has been designed, a unique set of 500 tuples have been used for testing procedure

<sup>1</sup> Existing benchmarks which have been implemented and available for NE type identification purposes.

considering the above-mentioned testing related hyper-parameter settings. As per the Table IV, the respective highest performance has been demonstrated once the value settings have been set to 0.6.

TABLE IV. RESULTS ON WORD EMBEDDING

Set	Dev Detect	Test Detect	Dev All	Test All
500	84.68	84.27	84.39	84.76

Once the word embeddings have been obtained, respective evaluation has been conducted considering the most prominent word embedding techniques such as Glove, Word2Vec, ELMo and BERT. The respective comparison can be exhibited under Fig. 7 as follows. As per the comparison when the k value gets 1.4 a performance increment can be observed considering all models except ELMo. When Word2Vec, Glove and BERT models have been considered, the highest outcome has been demonstrated by our approach which is the Glove model. Even though BERT has shown some promising results in most of the previous scenarios as per the constructive literature survey, the Glove model can be mentioned as an optimal alternative approach, especially for low resource language settings.

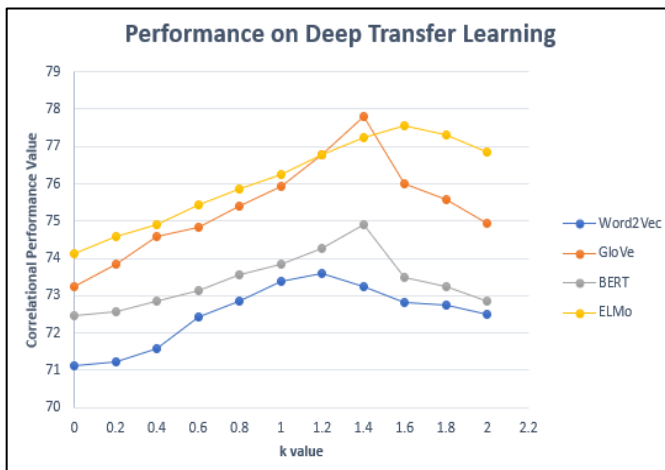


Fig. 7. Performance Comparison on Word Embedding through Transfer Learning.

Later, with the introduction of cross-lingual word embeddings, some state-of-the-art word embedding models have been invented. One of such prominent mechanisms would be the usage of XLM-R for obtaining word embeddings considering major text classification tasks [38] [39]. Hence, another evaluation approach has been applied to compare the proposed model and the XLM-R model where the results can be demonstrated under the Fig. 8 as follows.

As per the comparison between the GloVe and XLM-R models, it can be concluded that the maximum outcome has been showcased by GloVe even though XLM-R has performed well in most of the sections in the graph distribution.

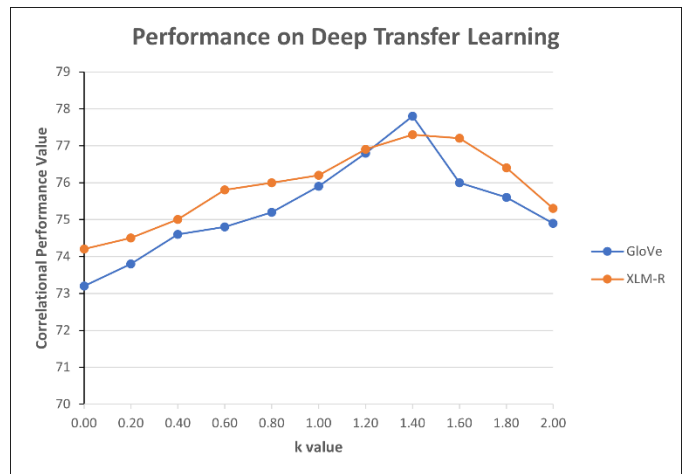


Fig. 8. GloVe and XLM-R Performance Comparison on Transfer Learning.

### C. Evaluation on NE Head Word Detection

NE head word detection has been identified as one of the scientific core novelties of this entire research attempt. Considering the major evaluation matrices which have been stated under the word embedding and the head word loss value, the testing procedure can be determined. Considering the key factors such as NE mention identifier and the head word detection loss value, the respective experimental results can be obtained and listed under Table V as follows.

As per the evaluation purposes, the existing NE head word detection mechanisms related to the flat NEs have been considered. Those existing benchmarks models have already been critically analyzed and reviewed under the comprehensive literature survey and the evaluation procedure has been conducted respectively. The respective evaluation results for the NE head word detection process can be showcased as per Table VI. Our novel model has been denoted as BB2022 (Boundary Bubbles). As per the competitive analysis, the most prominent deep neural based NE detection models such as Sohrab et al. 2018 [11], Ju et al. 2019 [16] and ARN2019 [15] have been used. These models have been recompiled and re-generated using the openly available source code in the respective code repositories. Considering the evaluation results it can be concluded that our novel model has performed well against all the prominent baselines which are available for NE head word detection.

TABLE V. RESULTS ON NE HEAD WORD DETECTION

Metric	Precision (%)	Recall (%)	F1 (%)
Dev Detect	85.973315	78.53125	82.083945
Test Detect	78.783828	81.76144	
Dev All	0.2187500	0.239808	
Test All	0.2338790	0.253577	
Dev Gap	85.707962	78.31259	81.844137
Test Gap	84.696074	78.54994	81.507865

TABLE VI. EVALUATION ON NE HEAD WORD DETECTION

Model	Set	Precision %	Recall %	F1 %
Sohrab2018	1	71.68	71.25	71.28
	2	71.59	70.48	71.24
	3	71.43	71.39	71.31
	4	72.69	71.59	71.84
	5	71.89	71.52	71.64
	6	72.98	71.42	71.55
Ju et al2019	1	69.11	68.87	68.91
	2	68.47	68.24	68.12
	3	68.35	68.29	68.23
	4	68.48	68.22	68.31
	5	68.94	68.47	68.51
	6	69.24	68.86	68.23
ARN2019	1	78.13	76.21	78.06
	2	78.21	76.47	78.17
	3	78.65	76.34	78.35
	4	77.98	75.27	77.76
	5	77.83	76.39	77.76
	6	78.11	77.53	78.10
BB2022	1	78.95	77.35	78.54
	2	78.24	76.97	78.13
	3	78.88	77.31	78.24
	4	78.65	78.03	78.59
	5	79.16	77.68	78.88
	6	79.14	77.21	78.92

#### D. Evaluation on NE Boundary Detection

This component has already been mentioned as the next scientific core novelty of this overall research procedure under the methodology section. In other words, this component would be the core driving force of the entire solution. The respective performance related matrices have been generated for the five most prominent NE categories and the respective results can be demonstrated under Table VII as follows.

TABLE VII. RESULTS ON NE BOUNDARY DETECTION

NE Type	Precision (%)	Recall (%)	F1 (%)
ORG	0.84	0.89	0.86
DESIG	0.85	0.85	0.85
LOC	0.96	0.96	0.96
PER	0.96	0.95	0.93
DATE	0.88	0.91	0.89

Additionally, the accuracy also has been calculated considering the major hyper-parameters. Several hyper-parameters have been set off such as word embedding rate to be 25, output dimension to be 4 and the dropout rate as 0.1. Once the results have been obtained, the evaluation has been conducted. For evaluation purposes, as per the benchmarks, the systems which are dedicated for detecting NE boundaries have been used. The obtained evaluation results can be tabulated under Table VIII as follows. Considering the evaluation-outcome it can be concluded that our novel model, BB2022 (Boundary Bubbles), has performed well than the existing baselines.

TABLE VIII. EVALUATION ON NE BOUNDARY DETECTION

Model	Precision %	Recall %	F1 %
Sohrab2018	76.61	69.20	72.71
Ju et al2019	79.90	67.08	71.36
ARN2019	81.34	68.20	72.83
<b>BB2022</b>	<b>82.18</b>	<b>71.34</b>	<b>76.54</b>

#### E. Evaluation on NE Linking

Conducting the evaluation on NE linking can be described as the final phase of the overall evaluation procedure. In terms of accomplishing the NE linking procedure, several major topics under the Sinhala speech knowledgebase have been used. As per the methodology, only the direct mapping technique of the NE linking has been used so that the direct mapping technique should be evaluated as the evaluation criteria on the overall NE linking. Considering the major NE linking benchmarks, several prominent systems such as DBpedia Spotlight<sup>2</sup>, SOTA<sup>3</sup> and VCG<sup>4</sup> can be listed. Our model has been evaluated with those existing benchmarks and the evaluation results can be showcased under Table IX as follows. As per the overall comparison, it can be concluded that our novel approach, BB2022 (Boundary Bubbles), has performed better than the existing benchmarks.

TABLE IX. EVALUATION ON NE LINKING

Model	Precision %	Recall %	F1 %
DBpedia Spotlight	52.64	52.48	52.53
SOTA	57.26	51.24	55.24
VCG	61.25	58.26	60.18
<b>BB2022</b>	<b>62.47</b>	<b>60.15</b>	<b>61.27</b>

#### V. DISCUSSION AND CONCLUSION

NE boundary detection in Sinhala context can be introduced as the core of this entire research attempt. The usage of social media for various different kinds of purposes has been increased in an alarming rate [8] [9]. There is a high demand for a sustainable computational framework for identifying and extracting such kind of various inputs which have been spread out in social media, especially in low NLP resources-based contexts like Sinhala [2]. Considering the overall approach of HelaNER 2.0, various enhancements have been designed, implemented, and evaluated. Major enhancements have been applied under the methodological components of word embedding through deep transfer learning, NE boundary detection and NE linking. All the predefined objectives such as Sinhala social media analysis, obtaining word embedding through transfer learning, NE head word detection, NE boundary detection and NE linking have been achieved. The novelties have been achieved in NE head word detection and NE boundary detection components. Respective evaluations have been conducted and the outcome has revealed that the novel approaches have outperformed the existing baselines in the market. Several points can be summarized as follows.

<sup>2</sup> <https://www.dbpedia-spotlight.org/api>

<sup>3</sup> Exploring Neural Entity Representations for Semantic Information, A Runge, E Hovy - arXiv preprint arXiv:2011.08951, 2020 - arxiv.org

<sup>4</sup> <https://paperswithcode.com/task/entity-linking>

Compared to the performance dimensions of the previous approaches, several avenues have been identified as the key performance indexes (KPIs) for HelaNER 2.0. Firstly, word embedding through deep transfer learning has been tuned-up in terms of considering character level feature representations. Secondly, NE head word detection has been improved by enhancing the NE head word detection loss function. Also, multiple MLP and CRF layers have been adapted as another major improvement. Several experimentations have been conducted based on different hyper-parameter value adjustments. Thirdly, NE boundary detection has been further enhanced by introducing stack pointer networks which can be identified as the core scientific contribution for the domain of NE boundary detection. A constructive evaluation process has been followed taking the most prominent approaches which have been established as baselines. The overall evaluation procedure has been strengthened further through increasing the overall evaluation cycles. As per the evaluation outcome, it can be concluded that our novel avenues which have been introduced under HelaNER 2.0, have outperformed the existing benchmarks. As it's been mentioned earlier, even though this approach can be mentioned as a niche specialized solution for Sinhala domain, the generalized version of this would make a real impact for NE boundary detection in other domains.

Considering all these aspects, few potential avenues can be introduced as possible future enhancements under the whole process of detecting Sinhala related NE boundaries along with the considered NE type. It has been assumed as a particular head word would not be shared among more than one sentence nuggets under the section of dominant word detection considering MLP on top of CRF along with NE type detection in the constructive methodological approach. Therefore, sharing a given respective head word among several multiple sentence nuggets has been considered as a possible future enhancement under the overall methodological approach. Experimenting on the applicability of sharing a particular head word among multiple different sentence nuggets would be a major scientific research advancement of this entire domain.

#### ACKNOWLEDGMENT

This research was supported by the Accelerating Higher Education Expansion and Development (AHEAD) Operation of the Ministry of Higher Education Sri Lanka funded by the World Bank.

#### REFERENCES

- [1] Jing Li, Aixin Sun and Yukun Ma, "Neural Named Entity Boundary Detection", IEEE Transactions on Knowledge and Data Engineering, 2015.
- [2] Y.H.P.P Priyadarshana, L. Ranathunga, C.R.J Amalraj and I. Perera, "HelaNER: A Novel Approach for Nested Named Entity Boundary Detection," presented at the IEEE 19th International Conference on Smart Technologies (EUROCON), Lviv, Ukraine, Jul. 6-8, 2021.
- [3] Abhishek Abhishek, Ashish Anand, and Amit Awekar, "Fine-Grained Entity Type Classification by Jointly Learning Representations and Label Embeddings", in Proc. of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, 2017.
- [4] Ioannis Partalas, Cédric Lopez, Nadia Derbas and Ruslan Kalitvianski, "Learning to Search for Recognizing Named Entities in Twitter", in Proc of the 2nd Workshop on Noisy User-generated Text (WNUT), 2016.
- [5] Jing Li, Deheng Ye and Shuo Shang, "Adversarial Transfer for Named Entity Boundary Detection with Pointer Networks", Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI-19), 2019.
- [6] Jason P.C. Chiu and Eric Nichols, "Named Entity Recognition with Bidirectional LSTM-CNNs", Transactions of the Association for Computational Linguistics, Volume 4, 2016.
- [7] R. Collobert et al., "Natural language processing (almost) from scratch," The Journal of Machine Learning Research, 12:2493–2537, 2011.
- [8] Changmeng Zheng, Yi Cai, Jingyun Xu, Ho-fung Leung and Guandong Xu, "A Boundary-aware Neural Model for Nested Named Entity Recognition", in Proc of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019.
- [9] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer, "Neural Architectures for Named Entity Recognition", in Proc of NAACL-HCT 2016, pp. 260-270.
- [10] Miguel Ballesteros, Chris Dyer, and Noah A. Smith, "Improved transition-based dependency parsing by modeling characters instead of words with LSTMs", In Proc of EMNLP, 2015.
- [11] Mohammad Golam Sohrab and Makoto Miwa, "Deep Exhaustive Model for Nested Named Entity Recognition", in Proc of the Conference on Empirical Methods in Natural Language Processing, 2018, pp. 2843-2849.
- [12] Guodong Zhou, Jie Zhang, Jian Su, Dan Shen, and Chewlim Tan, "Recognizing Names in Biomedical Texts: A Machine Learning Approach, Bioinformatics," 20(7):, 2004, pp.1178–1190.
- [13] Arzoo Katiyar and Claire Cardie, "Nested Named Entity Recognition Revisited," in Proc of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), New Orleans, Louisiana. ACL, 2018, pp. 861–871.
- [14] Nguyen Truong Son and Nguyen Le Minh, "Nested Named Entity Recognition Using Multilayer Recurrent Neural Networks," in Proc. of PACLING 2017, Sedona Hotel, Yangon, Myanmar, 2017, pp 16–18.
- [15] Hongyu Lin, Yaojie Lu, Xianpei Han and Le Sun, "Sequence-to-Nuggets: Nested Entity Mention Detection via Anchor-Region Networks," in Proc. of the 57th Annual Meeting of the Association for Computational Linguistics, 2019.
- [16] Wang, B., and Lu, W, "Neural segmental hypergraphs for overlapping mention recognition," In Proc. of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018, pp. 204–214.
- [17] Wang, B., Lu, W., Wang, Y., and Jin, H, "A neural transition-based model for nested mention recognition," in Proc. of the Conference on Empirical Methods in Natural Language Processing, 2018, pp. 1011–1017.
- [18] Xu, M., Jiang, H., and Watcharawittayakul, S, "A local detection approach for named entity recognition and mention detection," in Proc. of the 55th Annual Meeting of the Association for Computational Linguistics, volume 1, 2017, pp. 1237–1247.
- [19] Chuanqi Tan, Wei Qiu, Moshu Chen, Rui Wang and Fei Huang, "Boundary Enhanced Neural Span Classification for Nested Named Entity Recognition," The Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-20), 2020.
- [20] Xiaoya Li et al., "A Unified MRC Framework for Named Entity Recognition", in Proc. of the 58th Annual Meeting of the Association for Computational Linguistics, 2020.
- [21] Omer Levy, Minjoon Seo, Eunsol Choi, and Luke Zettlemoyer, "Zero-shot relation extraction via reading comprehension," in Proc. of the 21st Conference on Computational Natural Language Learning (CoNLL 2017), 2017, pp. 333–342.
- [22] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in Proc. of NAACL-HLT, 2019, pp. 4171–4186.

- [23] Y. Chen et al, "A Boundary Regression Model for Nested Named Entity Recognition," cited at Computation and Language (cs.CL); Artificial Intelligence (cs.AI) as arXiv:2011.14330 [cs.CL], 2020.
- [24] J. Sivic and A. Zisserman, "Video google: A text retrieval approach to object matching in videos." In null, 2003, pp. 1470.
- [25] Chao Zhang, Zichao Yang, Xiaodong He, and Li Deng. "Multimodal intelligence: Representation learning, information fusion, and applications." arXiv preprint arXiv:1911.03977, 2019.
- [26] H. M. S. T. Sandaruwan, S. A. S. Lorensuhewa and M. A. L. Kalyani, "Sinhala Hate Speech Detection in Social Media using Text Mining and Machine learning," 19th International Conference on Advances in ICT for Emerging Regions (ICTer), Colombo, Sri Lanka, 2019, doi: 10.1109/ICTer48817.2019.9023655, pp. 1-8.
- [27] Gitari, N. D. et al., "A lexicon-based approach for hate speech detection," International Journal of Multimedia and Ubiquitous Engineering, 10(4),. doi: 10.14257/ijmue.2015.10.4.21, 2015, pp. 215–230.
- [28] Cambria, E. et al., "SenticNet : A Publicly Available Semantic Resource for Opinion Mining," Artificial Intelligence, doi: 10.1038/leu.2012.122, 2010, pp. 14–18.
- [29] Köffer, S. et al., "Discussing the Value of Automatic HateSpeech Detection in Online Debates," Multikonferenz Wirtschaftsinformatik, October 2018. doi: 10.1111/j.1365-2923.2008.03277.x, pp. 83–94.
- [30] Alfina, I. et al., "Hate speech detection in the Indonesian language: A dataset and preliminary study," International Conference on Advanced Computer Science and Information Systems, ICACSIS 2017, 2018–October, doi: 10.1109/ICACSIS.2017.8355039, pp. 233–237.
- [31] W. Warner and J. Hirschberg, "Detecting hate speech on the world wide web," in Proc. of the Second Workshop on Language in Social Media, LSM '12, 2012, pp. 19-26.
- [32] C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad and Y. Chang, "Abusive language detection in online user content," in Proc. of the 25th International Conference on World Wide Web, WWW'16, 2016, pp. 145-153.
- [33] Z. Waseem and D. Hovy, "Hateful symbols or hateful people? predictive features for hate speech detection on twitter," in Proc. of NAACL-HLT, 2016, pp. 88-93.
- [34] P. Badjatiya, S. Gupta, M. Gupta and V. Varma, "Deep learning for hate speech detection in tweets," in Proc. of the 26th International Conference on World Wide Web Companion, 2017, pp. 759-760.
- [35] F. Vigna, A. Cimino, F. DellOrletta, M. Petrocchi and M. Tesconi, "Hate me, hate me not: Hate speech detection on facebook," in Proc. Of the First Italian Conference on Cybersecurity, 2017, pp. 86 – 95.
- [36] Y. Chen et al., "A Boundary Assembling Method for Nested Biomedical Named Entity Recognition," in IEEE Access, vol. 8, pp. 214141-214152, 2020, doi: 10.1109/ACCESS.2020.3040182.
- [37] Gwet, Kilem L. Large-Sample Variance of Fleiss Generalized Kappa, Educational and Psychological Measurement 81, 2021, pp. 781-790.
- [38] A. Alcoforado et al., ZeroBERTo: Leveraging Zero-Shot Text Classification by Topic Modeling, 15th International Conference on Computational Processing of Portuguese, 2017.
- [39] Efsun Sarioglu Kayi, Linyong Nan, Bohan Qu, Mona Diab, and Kathleen McKeown, Detecting Urgency Status of Crisis Tweets: A Transfer Learning Approach for Low Resource Languages, In Proceedings of the 28th International Conference on Computational Linguistics, 2020, pp. 4693–4703.

# Face Recognition under Illumination based on Optimized Neural Network

Napa Lakshmi<sup>1</sup>

Computer Science and Engineering  
Presidency University, Bangalore, India

Megha P Arakeri<sup>2</sup>

Information Science and Engineering  
Ramaiah Institute of Technology, Bangalore, India

**Abstract**—Face recognition is a significant area of pattern recognition and computer vision research. Illumination in face recognition is obvious yet challenging task in pattern matching. Recent researchers introduced machine learning algorithms to solve illumination problems in both indoor and outdoor scenarios. The major challenge in machine learning is the lack of classification accuracy. Thus, the novel Optimized Neural Network Algorithm (ONNA) is used to solve the aforementioned drawback. First, we propose a novel Weight Transfer Ideal Filter (WTIF) which is employed for pre-processing to remove the dark spots and shadows in an image by normalizing low frequency and high frequency of illumination. Secondly, Robust Principal Component Analysis (RPCA) is employed to perform efficient extraction of features based on image area representation. These features are given as input to ONNA which classifies the given input image under illumination. Thus we achieve the recognition of the face under various illumination conditions. Our approach is analyzed and compared with existing approaches such as Support Vector Machine (SVM) and Random Forest (RF). ONNA is better in terms of high accuracy and low error rate.

**Keywords**—Face recognition; illumination; neural network; robust principal component analysis

## I. INTRODUCTION

Face recognition is an interesting research area in computer science and information technology since 1990 [1] with several applications such as biometrics, law enforcement and surveillance video systems. Various face recognition methods have been developed during the previous two decades, and significant progress has been reached. Consequently, the efficiency of facial recognition systems under controlled conditions has already achieved a sufficient level. Unfortunately, the recognition rate of the existing FR system needs to be enhanced, also in real-world conditions like noise, illumination changes, pose and disguise. Because the accuracy of a face recognition technique varies a lot depending on the type of illumination and lighting used, illumination is one of the major aspects to consider when creating a human face recognition system. Due to the 3D geometry of human faces, it is noted that variations in illumination conditions form various shading and shadows on the face. This may make some facial characteristics appear weaker, or it may cause bright or dark areas in face images.

Face recognition accuracy is now quite high under perfect illumination; however, there are still issues to be solved under varying illumination [2, 3]. Several approaches have been developed in recent decades to improve the accuracy of facial

recognition in different illumination conditions, which can be divided into three groups: illumination modeling, illumination extraction of invariant features, and face image normalization [4]. The Retinex illumination model [5] describes the apparent illumination at every spot on the face in terms of that point's inherent reflectance, as well as the amount and angle of incident illumination. According to this concept, the low spatial frequency elements of the face image convey information on illumination, but the components with a high spatial frequency indicate inherent sensitivity of the face that ought to be retrieved for recognition.

As a result, self-quotient imaging [6] is used to estimate the illumination on a smoother version of the image of the face that reduced the logarithmic of the actual face image to produce an unchangeable representation. Similar filtering can be done in the frequency to normalize the low-frequency illumination and high-frequency illumination components using a logarithmic Gaussian Band Pass Bilateral Filter [7]. As previously established, the way faces are represented in the actual world is always flawed, as well as the images contain significant flaws. As a result of its extreme sensitivity to the flaws and inability to cope with information that is lacking, the traditional PCA estimation may be distant from the underlying real distribution of the facial image.

To address this issue, a significant sparse learning framework named Robust Principal Component Analysis (RPCA) is used to extract the features [8] [9]. Some modern technologies take advantage of the similarities that every human face possesses. To find the most matching area at each face location, the area representation is achieved using a 2D Fourier magnitude spectrum [10].

In this research, we propose a Weight Transfer Ideal Filter (WTIF) for face recognition that is robust to illumination variation. It is employed to remove dark spots, shadows and reflections in an image. Further, Robust Principal Component Analysis (RPCA) is used to extract features based on image area representation. The proposed WTIF method is used to recognize the face accurately based on a position-based voting scheme by increasing the features matching. The Optimized Neural Network Algorithm (ONNA) improves the classification accuracy under illumination and the proposed method is analyzed and compared with existing methods such as Support Vector Machine (SVM) and Random Forest (RF). Consequently, ONNA is better in terms of Accuracy, Equal Error Rate.



## II. RELATED WORK

Kim et al. [11] presented a method in which ground truth image could be employed to train the Illumination Normalization (IN) method using a convolutional neural network to convert the illumination variant face image into a feature map. The result showed that the IN-Net achieved better Face Recognition (FR) accuracy. Wu et al. [12] discussed the face recognition method across the posture and illumination issue, given a modest collection of training samples and one sample per gallery. The different illumination samples and deep neural network capacity to perform nonlinear transformations make multitask deep neural network ideal for posture and illumination normalization. Results show that this algorithm achieved better results on MultiPIE dataset less training data and also verified the effectiveness of introduced method.

Xiangpo Wei et al. [13] developed a face recognition approach based on a convolutional neural network (CNN) and a local binary pattern feature extraction method (LBP) that overcomes the impacts of illumination, posture, and expression. Local Binary Pattern (LBP) represents the local texture features of an image. CNN is capable of extracting image features and reducing their dimensionality. The experimental result demonstrated that the method could significantly increase the rate of accuracy and also has better robustness of illumination and posture.

Khan et al. [14] presented Particle Swarm Optimization (PSO) to optimize textural and wavelet features. The Discrete Wavelet Transform had the benefit of isolating significant characteristics, which reduced processing time and improved recognition accuracy. The results showed that the suggested technique is superior and resistant to illumination and has good accuracy rate. Han et al. [15] discussed the Accelerated Proximal Gradient (APG) technique and illumination regression filtering that was applied to remove the illumination effect. The results showed the introduced technique was robust to illumination.

Liang et al. [16] presented a method where Wavelet transform images of the Low Frequency Sub-Band (LFSB) and High Frequency Sub-Band (HFSB) were boosted and denoised, which frequently resulted in a loss of information due to less attention of HFSB. Furthermore, the Particle Swarm Optimization-Neural Network (PSO-NN) was used to classify the data. The suggested network can effectively create a robust visual impact under varied illumination and greatly increase recognition performance, according to experimental results.

Zhang et al. [17] presented the Expected Patch Log Likelihood (EPLL) algorithm that extracted illumination weight and used Neighboring Radiance Ratio algorithm (NRR) which optimized the initial vector of the Gaussian mixture model that utilized redundant data in image. Dewantara et al. [18] presented a novel optimized fuzzy based illumination constant approach that overcomes the influence of illumination for photometric based face recognition. It efficiently eliminates the impact of illumination on facial images and has a high level of robustness. Output proved that the introduced algorithm improved computational time and also that improve the face detection performance.

Baradarani et al. [19] presented the multi resolution analysis and sub band filtering the Double-Density Dual-Tree Complex Wavelet Transform (DD-DTCWT) was helpful and easy for illumination invariant face recognition. Principal Component Analysis (PCA) was employed for feature extraction and the Extreme Learning Machine (ELM) was used for faster classification. The result proved that the introduced method has high recognition rate and computational complexity.

Vidya et al. [20] used Discrete Wavelet Transform (DWT) that aided in the efficient extraction of features and the introduced Selective Illumination Enhancement Technique (SIET) which has high incidence of recognition. The result showed that the introduced method has average recognition rate for Color FERET database. Yang et al. [21] presented Nuclear Norm based Matrix Regression (NMR) for face recognition and categorization. Result showed that the NMR was more reliable for recognition with illumination changes.

## III. PROPOSED METHODOLOGY

Face recognition consists of three sections such as preprocessing, feature extraction, and classification. Fig. 1 explains the proposed methodology. Optimized Neural Network Algorithm (ONNA) is used to improve the classification accuracy under illumination conditions and also recognizes the face accurately. The scheme comprises preprocessing using the proposed method Weight Transfer Ideal Filter, to find the most matching area at each face location using a 2-D Fourier magnitude spectrum. Robust Principal Component Analysis (RPCA) is used for feature extraction. ONNA along RPCA gave better results in terms of high accuracy and low error rate.

### A. Weight Transfer Ideal Filter

The main purpose of pre-processing is to improve image quality so that it can be processed further by removing or minimising unrelated and surplus components in the illumination images. Weight Transfer Ideal Filter (WTIF) is employed to remove dark spots, shadows and reflection and also it reduces low frequency illumination and high frequency illumination. WTIF technique is used to recognize the face accurately based on a position-based voting scheme by increasing the features matching. A Weight Transfer Ideal Filter is a complete filter that combines spatial domain and range domain filtering to remove noise while preserving edge characteristics using (1) and (2).

$$\mu = \frac{1}{ab} \sum_{n_1 n_2} m(n_1, n_2) \quad (1)$$

$$Z_i = \sum_{k=0}^i P_{i,j} \cdot Q_{i,j} \cdot y_i \quad (2)$$

In Weight Transfer Ideal Filter, weights of the pixel decay as a function of distance from the filter's center, as provided by,

$$G_\sigma(a, b) = \frac{1}{2\pi\sigma^2} e^{-\frac{(a^2+b^2)}{2\sigma^2}} \quad (3)$$

$$I_F(B) = \frac{1}{W} \sum_{q \in S} G_{\sigma_s}(\|P - Q\|) G_{\sigma_r}(|I(P) - I(Q)|) I(Q) \quad (4)$$

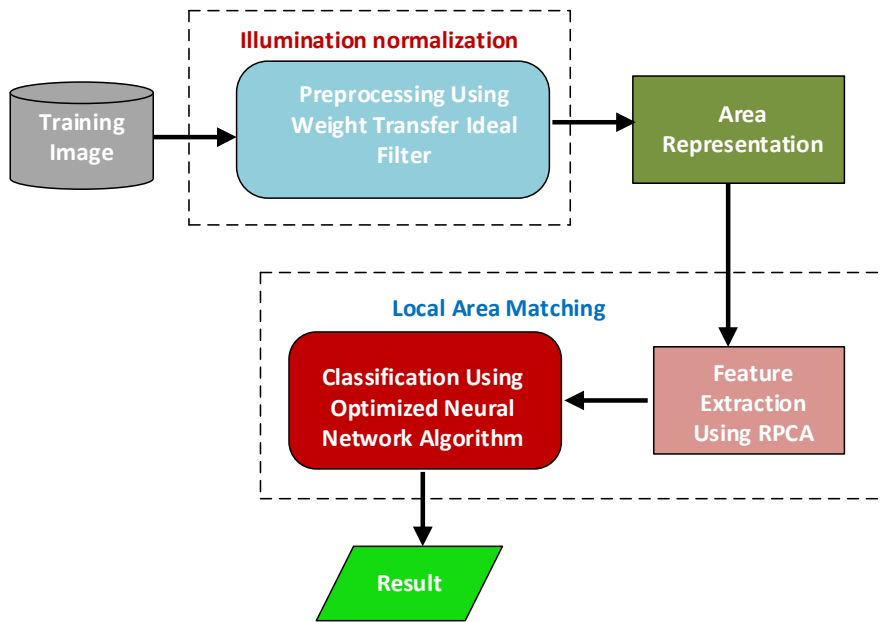


Fig. 1. Proposed Methodology.

The Weight Transfer Ideal Filter is used for smoothing nonresponsive noise from two-dimensional signals while maintaining image quality. As a result, it's particularly well-suited to improve the illumination of the face image. The pre-processing technique is utilized in illumination improvement, artifact removal, and alignment. The pre-processing technique involved, creates masks for pixels with the greatest amount of effort to reduce dark spots and reflection and Fourier Magnitude Spectra as Image Area Representation Features.

Due to slight changes in a person's facial features and head position, even with properly matched face images, matching local pixel areas on the faces may not correlate to the same spatial structure. We employ a 2-D Fourier magnitude spectrum as the feature to describe each local image area to limit the outcome of modest mismatch errors in recognition [23]. We use the shift-invariance of the Fourier magnitude representation to enhance the ability to withstand minor mismatch errors and variations in face expression by taking the magnitude spectrum instead of the phase spectrum. When the 2-D Fourier transform is applied to an area in a Nonlinear Active Band pass Filter testing image, the magnitude spectrum of the resulting testing image may be represented as

$$|\tilde{I}_{\delta(a,b)}(m, n)| \cong \beta_{\delta(a,b)} |\widetilde{R}_{\delta(a,b)}(m, n)| + \delta(m, n) \alpha_{\delta(a,b)} \quad (5)$$

Where  $|\tilde{I}_{\delta(a,b)}(m, n)|$  and  $|\widetilde{R}_{\delta(a,b)}(m, n)|$  indicate the spectral magnitudes of the Gaussian band pass bilateral filter pixel value and intrinsic reflectivity of the area,  $\alpha_{\delta(a,b)}$  indicate the residual constant illumination background and the  $\delta(m, n)$  is the Kronecker delta function.

### B. Feature Extraction using Robust Principal Component Analysis

The output of the matching area representation is given as the input to the feature extraction. The Robust Principal Component Analysis (RPCA) method is used to extract textural features accurately by reducing the sparse error.

Consider a huge data matrix  $H \in \mathbb{R}^{m \times n}$  has a reduced layout  $I$  but is contaminated by sparse errors element  $M$ , i.e.,  $H=I+M$ . The goal is to reclaim a low-rank element  $I$  from the substantially corrupted matrix  $H$  in a reliable manner. Unlike in conventional PCA, the noise is small, the entries in  $M$  might have any magnitude and their support is believed to be simple but uncertain. The following is the original concept of Robust Principal Component Analysis (RPCA) [24]:

$$\min (\text{rank} (I)+ \gamma \|M\|_0), \quad \text{s.t. } H=I+M \quad (6)$$

Where  $\|M\|_0$  represents the matrix  $\ell_0$  norm that is collecting nonzero components in the matrix  $H$ . Due to the rank measure's non-smoothness and non-convexity as well as the zero-norm penalty, (1) is hard to solve. Principal Component Pursuit (PCP) is solved in the relaxed form using tractable convex optimization:

$$\min (\|I\|_* + \gamma \|M\|_1), \text{ s.t. } H=I+M \quad (7)$$

Where the rank procedure in (6) is nuclear matrix has taken its place  $\|\cdot\|_*$ , the matrix  $\ell_1$  -norm estimates the matrix  $\ell_0$ -norm and  $\gamma$  is regulation variable for balance and has been fixed to  $1/\sqrt{\max(m, n)}$ . It has been demonstrated both

mathematically and practically that the resolution of (7) properly retrieves the low-rank and sparse elements under very weak conditions, as provided as the rank of  $I$  is not too great and the errors  $M$  is sparsely maintained [22].

### C. Optimized Neural Network (ONNA)

An Optimized neural network is used to improve the classification accuracy under illumination. Firstly, search space algorithm is used to extract the features of the image. A search in a search space method is an example of a possible solution to the problem of determining the relevance of each characteristic. Suppose there are  $n$  search  $k$  dimensional space, then the position of search  $i$  can be represented as  $X_i = (x_{i,1},$

$x_{i,2}, \dots, x_{i,k}$  ( $i=0,1,2, \dots, n$ ). The velocity and position of each search are updated as follows:

$$V_i(t+1) = \omega V_i(t) + C_1 r_1 [pbest(t) - x_i(t)] + C_2 r_2 [gbest(t) - x_i(t)] \quad (8)$$

$$x_i(t+1) = x_i(t) + V_i(t+1) \quad (9)$$

Where  $C_1$  is the cognitive coefficient and  $C_2$  is the social coefficient and matching values  $r_1$  and  $r_2$  are two independently generated random numbers and  $\omega$  is the inertia weight. The maximum generations or the better position of the object in the cluster is no more included in the search algorithm, which cannot be improved even after a many number of generations. Therefore, the proposed searching algorithm has proved its efficiency and robustness in overcoming complex optimization challenges.

The range of  $C_1$  and  $C_2$  are  $C_1 \in (2.75, 1.25)$  and  $C_2 \in (0.5, 2.25)$  respectively. The learning factor function expression of linear change is described as given below using (10) and (11).

$$\Delta Pbest = g_1 * rand(0,1) * (Pbest_{i,j} - x_{i,j}) \quad (10)$$

$$\Delta Gbest = g_2 * rand(0,1) * (Gbest_{i,j} - x_{i,j}) \quad (11)$$

Where  $C_1$  and  $C_2$  are learning factors; rand is a positive random number between 0 and 1 distributed normally. This search algorithm is mainly used to learn the features. Given the face images  $q_1, q_2, \dots, q_m$ , the average face of these given face images is defined by (12).

$$\varphi = \frac{1}{s} \sum_{i=1}^s q_i \quad (12)$$

The difference between each input face and the average face is calculated as follows

$$\Psi_i = q_i - \varphi \quad (13)$$

The covariance matrix CM can then be computed using

$$CM = \sum_{i=1}^N \Psi_i \Psi_i^T = AA^T \quad (14)$$

The component effective method is used to train the learned sample. Finally, the error-reduced training sample is sent to the classification phase; here the neural network is issued to classify the training sample. It classifies the face accurately and finalizes feature matching and recognition. Finally, the feature matrices that correlate to numerous facial images are sent to the neural network as training data. The neural network driven feature learning framework, consists of 3 layers i.e., input layers, an output layer, and a hidden layer. For the Facial illumination Recognition problem, the right multiplication projection matrices in the hidden layer are also employed to get more discriminative characteristics among the high-level characteristics.

Let  $G_i = \{G_{i,j}^t, j = 1, \dots, E_i\}$  ( $t = 1, \dots, P_t$ ) indicate the  $t^{th}$  multi-channel projection matrix set consisting of  $C_i$  channels of projection matrices, where  $G_{i,j}^{(t)}$  denote the  $j$  th channel matrix of  $G_i$ ,  $P_i$  indicate the number of multi-channel projection matrix sets, and  $E_t$  indicate the number of channel matrices in

$G_i$ . The hidden layer can therefore be represented in the following way:

$$C_i = \sum_{j=1}^{E_t} G_{i,j}^{(t)} Q_j, (i=1, 2, \dots, P_t) \quad (15)$$

Where  $C_i$  indicates the matrix in  $t^{th}$  channel of the output and  $Q_j$  is the  $j$ th channel of the input matrices. If the input layer performs proper multiplication,

$$C_i = \sum_{j=1}^{E_a} Q_j G_{i,j}^{(a)}, (t=1, 2, \dots, P_a) \quad (16)$$

Then each row of the input matrix is projected onto a different feature area, resulting in a dimension reduction result at the same time. The resulting error by the given mode or result, i.e., the weight of the present related edge in the network is computed by combining the hidden layer and output layer results. The error of the hidden neuron is also calculated using the correlation weight and the error of the next layer's neuron. The network weight is updated with each neuron's error computed. Finally, the optimized neural network algorithm classifies the images. Fig. 2 shows the steps in ONNA.

---

### Algorithm steps

---

Input: Dataset, image size  $P$ , inertia feature  $\omega$ , cognitive coefficient  $c1$ , social coefficient  $c2$ , and matching value

Step 1: Initialize the parameters  $x_i$ , pbest, and gbest

Step 2: Determine the best face feature value selection

Step 3: Determine and pbest and gbest value

Step 4: Update  $x_i$  for each feature using (17)

$$V_i(t+1) = \omega V_i(t) + C_1 r_1 [pbest(t) - x_i(t)] + C_2 r_2 [gbest(t) - x_i(t)] \quad (17)$$

$$x_i(t+1) = x_i(t) + V_i(t+1) \quad (18)$$

Step 5: If the matching value is not met, go to step 2

Step 6: Steps 5 and 6 should be repeated until the set minimal error or the matching value is reached.

Step 7: Take the global best value of the training process.

Step 8: Set the learning rate, maximum iterations, number of hidden layer neurons, features and thresholds between the input and hidden layers, between the hidden and output layers, and the algorithm termination minimal allowable error.

Step 9: The training samples should be normalized.

Step 10: Start the training process by dividing the images into sub-images to match with the corrupted images and enhance the classification decision.

Step 11: Validate the original data by the trained process, and restores the output data to the original order of image.

Step 12: Testing Process

Step 13: Hybrid Weight Transfer Ideal Filter is used to recognize the face under illumination and matches the features. After that, the neural network is used to classify the image and implement the Optimized Neural Network Algorithm (ONNA) techniques to improve the classification.

Step 14: In this manner, a deep neural network classifier based on a position-based voting scheme is used to recognize the face accurately.

---

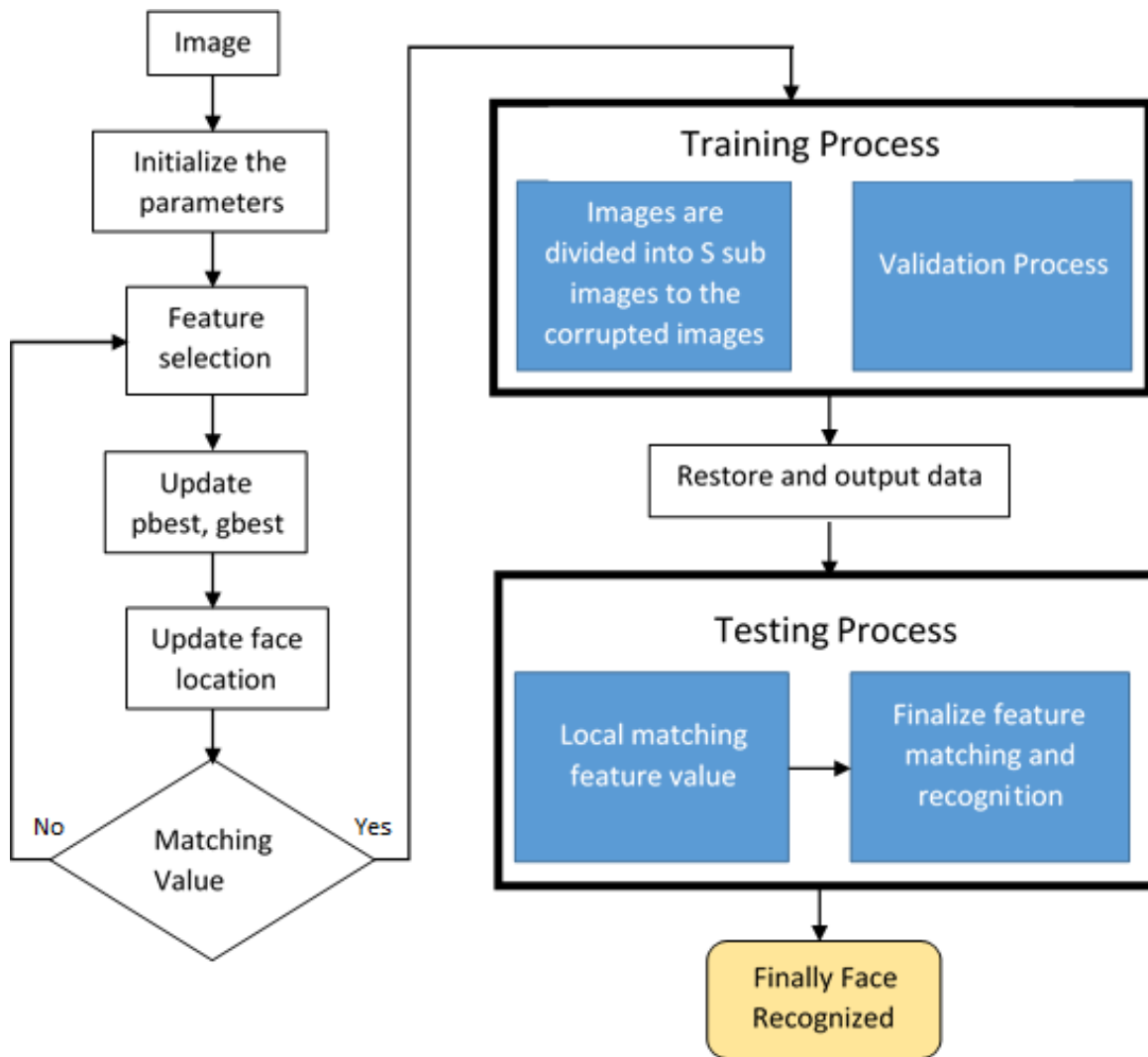


Fig. 2. Steps in Optimized Neural Network Algorithm.

#### IV. RESULTS AND DISCUSSIONS

All the experiments are performed in the Extended Yale-B dataset [25] and CMU Multi-PIE dataset [26]. Extended Yale-B database consists of 16128 images of 28 people under 64 different illumination conditions. CMU Multi-PIE database consists of 750,000 images of 337 people taken under 19 different illumination conditions. The Optimized neural network is introduced in this paper. Compared with SVM and RF the proposed method achieves high accuracy and low error rate.

The first column in Fig. 3 shows the input images full of dark spots. The proposed Weight Transfer Ideal Filter method enhances the low frequency images by removing the dark spots as shown in column 2. Column 3 images show the most matching area at each face location and the area representation is achieved using a 2-D Fourier magnitude spectrum. RPCA algorithm is applied to the images in column 3 to extract the features. Then the extracted features are given as an input to

the Optimized Neural Network Algorithm which improves the classification accuracy and by increasing local area matching, this method recognizes the face under illumination conditions. The recognized faces are shown in column 5.

The proposed method is compared with the existing classification methods such as support vector machine (SVM) and Random Forest (RF) in terms of accuracy and Error Rate as shown in Table 1 and Table 2. The accuracy of the proposed method is high compared to the existing classification methods because of the optimized neural network. Table 1 shows the results of face recognition under varied illumination on the Extended Yale-B database and Table 2 shows the results face recognition under varied illumination on the CMU-PIE database.

To prove the success rate of the introduced technique, it is essential to compare the Optimized Neural Network Algorithm with SVM and RF method, and the results are shown in Fig. 4.



Fig. 3. Dark Spot Removal and their Result.

TABLE I. PERFORMANCE ANALYSIS OF FACE ILLUMINATION IMAGE FOR OPTIMIZED NEURAL NETWORK ALGORITHM AND EXISTING METHOD ON EXTENDED YALE –B DATABASE

S.No	Performance analysis	SVM	RF	Proposed method
1	Accuracy	0.92	0.88	0.95
2	Error rate	0.037	0.074	0.03

TABLE II. PERFORMANCE ANALYSIS OF FACE ILLUMINATION IMAGE FOR OPTIMIZED NEURAL NETWORK ALGORITHM AND EXISTING METHOD ON CMU-PIE DATABASE

S.No	Performance analysis	SVM	RF	Proposed method
1	Accuracy	0.935	0.902	0.97
2	Error Rate	0.035	0.063	0.028

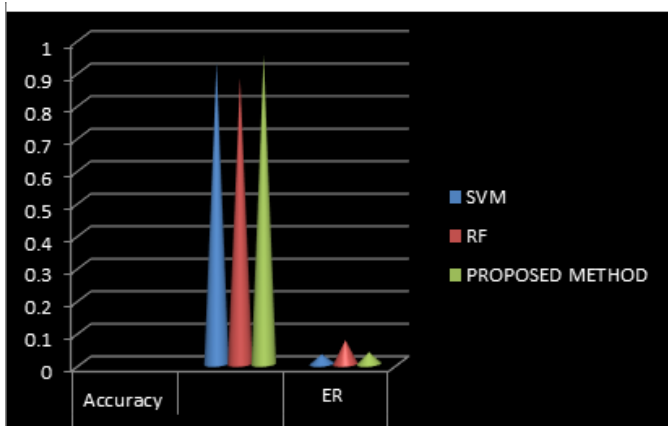


Fig. 4. Graphical Analysis of Optimized Neural Network Algorithm and Existing Classifier on Extended Yale-B Database.

Finally, the results showed that the performance of the proposed system is high in comparison to the existing methods.

## V. CONCLUSION AND FUTURE WORK

In the presence of varied illumination conditions, recognizing a face with good accuracy is the main challenge. In this paper, the Optimized Neural Network Algorithm (ONNA) is proposed to improve classification accuracy in varied illumination. Firstly, a pre-processing technique Weight Transfer Ideal Filter is proposed to reduce the dark spots, shadows, and reflection in the input image. Secondly, Robust Principal Component Analysis (RPCA) is applied to extract efficient features based on image area representation and the output of the RPCA is given as an input to the optimized neural network algorithm (ONNA) which improves the classification accuracy under illumination. Experiments are conducted on Extended Yale B and CMU-PIE datasets. According to the findings of the experiments, the ONNA outperforms the existing method such as SVM and RF in recognizing the faces under varied illumination in terms of high accuracy and low error rate. The future work is to use deep neural architectures such as Siamese Neural Network to improve the recognition rate.

## REFERENCES

- [1] Kim, Y-H., H. Kim, S-W. Kim, H-Y. Kim, and S-J. Ko. "Illumination normalisation using convolutional neural network with application to face recognition." *Electronics letters* 53, vol.6, pp. 399-401, 2017
- [2] Wu, Zhongjun, and Weihong Deng. "One-shot deep neural network for pose and illumination normalization face recognition." In 2016 IEEE International Conference on Multimedia and Expo (ICME), pp. 1-6. IEEE, 2016.
- [3] Ke, Pengfei, Maoguo Cai, Hanmo Wang, and Jialong Chen. "A novel face recognition algorithm based on the combination of LBP and CNN." In 2018 14th IEEE International Conference on Signal Processing (ICSP), pp. 539-543. IEEE, 2018.
- [4] Khan, Sajid Ali, Muhammad Ishtiaq, Muhammad Nazir, and Muhammad Shaheen. "Face recognition under varying expressions and illumination using particle swarm optimization." *Journal of computational science* vol. 28, pp. 94-100, 2018
- [5] Han, Xianjun, Yanli Liu, Hongyu Yang, Guanyu Xing, and Yanci Zhang. "Normalization of face illumination with photorealistic texture

- via deep image prior synthesis." *Neurocomputing*, vol.386, pp. 305-316, 2020.
- [6] Liang, Hongtao, Jie Gao, and Ning Qiang. "A novel framework based on wavelet transform and principal component for face recognition under varying illumination." *Applied Intelligence*, vol.51, issue 3 pp.1762-1783, 2021.
- [7] Zhang, Zijian, and Min Yao. "Illumination Invariant Face Recognition By Expected Patch Log Likelihood." In *SoutheastCon 2021*, pp. 1-4. IEEE, 2021.
- [8] Dewantara, Bima Sena Bayu, and Jun Miura. "OptiFuzz: a robust illumination invariant face recognition system and its implementation." *Machine Vision and Applications*, vol.27, issue 6, pp.877-891, 2016.
- [9] Baradarani, Aryaz, QM Jonathan Wu, and Majid Ahmadi. "An efficient illumination invariant face recognition framework via illumination enhancement and DD-DTCWT filtering." *Pattern Recognition*, vol.46, issue 1, pp.57-72, 2013.
- [10] Vidya, V., Nazia Farheen, K. Manikantan, and S. Ramachandran. "Face recognition using threshold based DWT feature extraction and selective illumination enhancement technique." *Procedia Technology*, vol.6, pp. 334-343, 2012.
- [11] Yang, Jian, Lei Luo, Jianjun Qian, Ying Tai, Fanlong Zhang, and Yong Xu. "Nuclear norm based matrix regression with applications to face recognition with occlusion and illumination changes." *IEEE transactions on pattern analysis and machine intelligence*, vol.39, issue 1, pp.156-171, 2016.
- [12] W.Zhao,R.Chellappa,P.J.Phillips,A.Rosenfeld, "Face recognition:a literature survey", *ACMComput.Surv*, 2013, pp.399-458.
- [13] Xiangpo Wei, Xuchu Yu, Bing Liu & Lu Zhi, "Convolutional neural networks and local binary patterns for hyperspectral image classification", *European Journal of Remote Sensing*, vol. 52, issue1, pp.448-462, 2019.
- [14] Wang JW, Le NT, Lee JS, Wang CC, "Illumination compensation for face recognition using adaptive singular value decomposition in the wavelet domain", *Information Sci*, vol.435, pp.69-93, 2018.
- [15] Nabatchian, Amirhosein, Esam Abdel-Raheem, and Majid Ahmadi. "Illumination invariant feature extraction and mutual-information-based local matching for face recognition under illumination variation and occlusion.", *Pattern Recognition*, vol.44, pp. 2576-2587, 2011.
- [16] H. Wang, S. Z. Li, and Y. Wang, "Face recognition under varying lighting conditions using self quotient image," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit.*, Seoul, South Korea, May 2004, pp. 819-824.
- [17] S. Srisuk and A. Petpon, "A Gabor quotient image for face recognition under varying illumination," in *Proc. 4th Int. Symp. Adv. Vis. Comput.*, Las Vegas, NV, USA, 2008, pp. 511-520.
- [18] Yang, Hong-Ying, Xiang-Yang Wang, Tian-Xiang Qu, and Zhong-Kai Fu. "Image denoising using bilateral filter and Gaussian scale mixtures in shiftable complex directional pyramid domain." *Computers & Electrical Engineering*, vol.37, issue 5, pp. 656-668, 2011.
- [19] E.Candes, X.Li, Y.Ma, J.Wright,"Robust principal component analysis" *Journal of the ACM*, vol.58(3), pp. 11:1-11:37, 2011.
- [20] J.Wright, A.Ganesh, S.Rao, Y.Ma, "Robust principal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization", *Neural Information Processing Systems*, 2009.
- [21] Y. Xu et al., "Data uncertainty in face recognition," *IEEE Trans. Cybern.*, vol. 44, no. 10, pp. 1950-1961, Oct. 2014.
- [22] Cai, Lian, and Sidan Du. "Rotation, scale and translation invariant image watermarking using Radon transform and Fourier transform." In *Proceedings of the IEEE 6th Circuits and Systems Symposium on Emerging Technologies: Frontiers of Mobile and Wireless Communication (IEEE Cat. No. 04EX710)*, vol. 1, pp. 281-284, 2004.
- [23] McLaughlin. N, Ming. J, Crookes. D. "Largest Matching Areas for Illumination and Occlusion Robust Face Recognition", *IEEE Transactions on Cybernetics*, vol. 47, issue 3, pp. 796-808, 2017.
- [24] Luan, Xiao, Bin Fang, Linghui Liu, Weibin Yang, and Jiye Qian. "Extracting sparse error of robust PCA for face recognition in the presence of varying illumination and occlusion", *Pattern Recognition*, vol. 47, issue 2, pp. 495-508, 2014.
- [25] A. Georghiades, et Al. "From few to many: illumination cone models for face recognition under variable lighting and pose." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(6):pp.643-660, 2001.
- [26] Sim, Terence & Baker, Simon & Bsat, Maan. (2002), "The CMU Pose, Illumination, and Expression (PIE) database", *Proceedings of the 5th IEEE International Conference*, 46-51. 10.1109/AFGR.2002.1004130.

# Transfer Learning for Medicinal Plant Leaves Recognition: A Comparison with and without a Fine-Tuning Strategy

Vina Ayumi<sup>1</sup>

Doctor of Engineering  
Universitas Sriwijaya, Indonesia

Handrie Noprisson<sup>4</sup>, Yuwan Jumaryadi<sup>5</sup>

Faculty of Computer Science  
Universitas Mercu Buana, Indonesia

Ermatita Ermatita<sup>2</sup>, Abdiansah Abdiansah<sup>3</sup>

Faculty of Computer Science  
Universitas Sriwijaya, Indonesia

Mariana Purba<sup>6</sup>

Program of Informatics  
Universitas Sjakhyakirti, Indonesia

Marissa Utami<sup>7</sup>, Erwin Dwika Putra<sup>8</sup>

Faculty of Engineering  
Universitas Muhammadiyah Bengkulu, Indonesia

**Abstract**—Plant leaves are another common source of information for determining plant species. According to the dataset that has been collected, we propose transfer learning models VGG16, VGG19, and MobileNetV2 to examine the distinguishing features to identify medicinal plant leaves. We also improved algorithm using fine-tuning strategy and analyzed a comparison with and without a fine-tuning strategy to transfer learning models performance. Several protocols or steps were used to conduct this study, including data collection, data preparation, feature extraction, classification, and evaluation. The distribution of training and validation data is 80% for training data and 20% for validation data, with 1500 images of thirty species. The testing data consisted of a total of 43 images of 30 species. Each species class consists of 1-3 images. With a validation accuracy of 96.02 percent, MobileNetV2 with fine-tuning had the best validation accuracy. MobileNetV2 with fine-tuning also had the best testing accuracy of 81.82%.

**Keywords**—Medicinal leaf plant; transfer learning; deep learning; phytomedicine

## I. INTRODUCTION

Leaves have characteristics such as shape and texture to be identified with the help of image processing technology and deep learning. An object sees identification as geometric information with boundaries [1]–[10]. The identified leaf object is limited to the boundary identified as leaf size and leaf shape, while the leaf texture or pattern is seen from the leaf surface [11]. Generally, the size of the leaves can be different, but the surface pattern of the leaves does not differ from one another [12]–[15]. This study aims to identify medicinal or phytomedicine plant species by processing leaf imagery using image processing and deep learning [15]–[21].

Research on the identification of phytomedicine plant leaves has been carried out by several previous studies, for

example Naresh and Nagendraswamy in 2015 [22], Mukherjee and his team in 2016 [23], Venkataraman & Mangayarkarasi in 2017 [24], Gao & Lin in 2018 [25], Sivaranjani et al. in 2019 [26], Pechebovicz et al. in 2020 [27], Bhuiyan et al. in 2021 [28].

In a study by Naresh and Nagendraswamy in 2015, the authors employed local binary patterns (LBP) to classify medicinal leaf plants. In a study conducted in 2016 by Mukherjee and his team, the classification of medicinal plants was accomplished with the use of Back Propagation Multi-Layer Perceptron (BP-MLP) [22], [29].

A study on the classification of medicinal plants also was carried out by Venkataraman and Mangayarkarasi (2017). They utilized the Histogram of Oriented Gradient (HoG)-Support Vector Machine for their research (SVM) [24]. Moreover, Gao and Lin (2018) used the OTSU approach in their classification of leaf plants for medicinal purposes. The OTSU approach involves using each manually marked edge point of the leaf to precisely detect the following outside points of the leaf located next to it [25]. The ExG-ExR index and the Logistic Regression (LR) classifier were utilized by Sivaranjani et al. to classify medicinal plants, and the researchers discovered that this method was successful. Based on each extracted leaf's color and texture characteristics, the Logistic Regression (LR) classifier is utilized to classify the various plant species [26].

In this study, we propose transfer learning models VGG16, VGG19, and MobileNetV2 to study the distinguishing features to identify medicinal plant leaves according to the dataset that has been collected. We also improved algorithm using fine-tuning strategy and analysed a comparison with and without a fine-tuning strategy to transfer learning model's performance.



## II. RELATED WORKS

Research on leaf image classification has been carried out for the last few years. To see the development of research in this field, we conducted a literature study on leaf image classification research from 2015 to 2021. The overview of related works is depicted in Fig. 1.

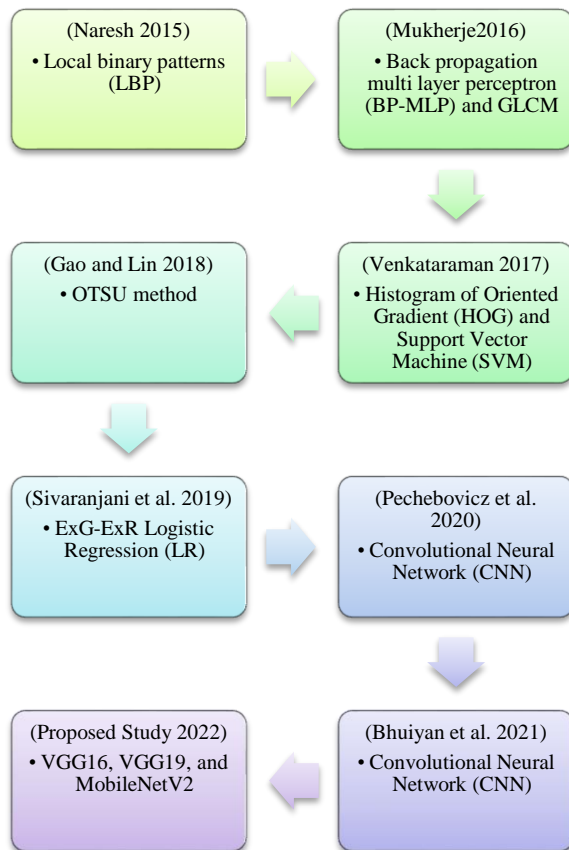


Fig. 1. Related Works

Local binary patterns (LBP) were used to classify medicinal leaf plants in a study by Naresh and Nagendraswamy in 2015 [22]. Back Propagation Multi-Layer Perceptron (BP-MLP) and Gray Level Co-occurrence Matrix (GLCM) were used to classify medicinal plants in a study by Mukherjee and his team in 2016. Results show that combined GLCM features can classify things better than basic single GLCM features. [23].

Venkataraman & Mangayarkarasi (2017) conducted a study for classification of medicinal plants using the Histogram of Oriented Gradient (*HoG*)-Support Vector Machine (SVM) [24]. Gao & Lin (2018) used the OTSU method to classify leaf plants that are used for medicine. OTSU method is to use each manually marked edge point of the leaf to accurately detect the next outer points of the leaf next to it [25].

Sivaranjani et al. (2019) used the ExG-ExR index and the Logistic Regression (LR) classifier to classify medicinal plants, and they found that this worked well. The Logistic Regression classifier is used to classify the plant species based

on the color and texture features of each extracted leaf. This classifier has a 93.3 percent accuracy rate [26]. Convolutional Neural Networks were used in a study by Pechebovicz et al. (2020) to classify medicinal plants [27]. Bhuiyan et al. (2021) used Convolutional Neural Networks to conduct research for the identification of medicinal plants (CNN) [28].

## III. RESEARCH METHODOLOGY

This research was conducted by applying several protocols or stages, including data collection, data preparation, feature extraction, classification, testing, and evaluation, as shown in the Fig. 2.

The first phase is data collection. The dataset used in this research is a public dataset called Medicinal Leaf Dataset. This dataset will be made public by Roopashree & Anitha in 2020 [30], [31]. The dataset collected is the result of photos using the Samsung s9+ Model Camera and Canon Inkjet Printer. Leaf photos are from leaves picked from different plants of the same species available at the study site. Healthy and mature leaves were selected for the dataset. The dataset consists of 1500 images of thirty species. Each species consists of 60 to 100 high quality images. An example of a dataset can be seen in the Fig. 3.

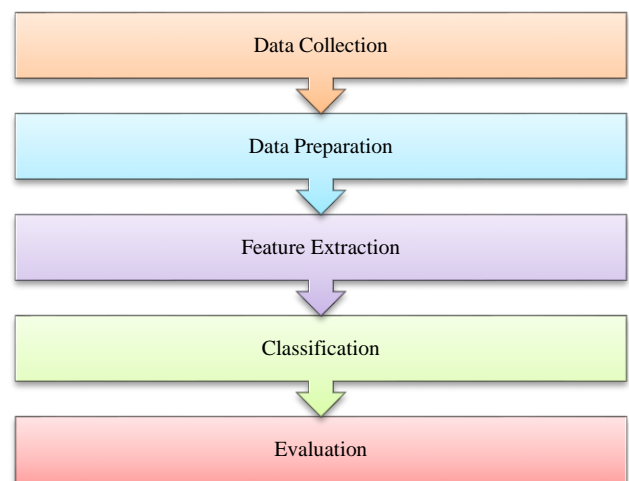


Fig. 2. Research Phases.



Fig. 3. Example of Dataset.

The dataset consists of 30 species of healthy medicinal plants such as *Alpinia Galanga* (Galanga Leaves), *Amaranthus Viridis* (Green Spinach Leaves), *Artocarpus Heterophyllus* (Jackfruit), *Azadirachta Indica* (Neem), *Santalum Album* (Sandalwood), *Muntingia Calabura* (Jamaica cherry), *Plectranthus amboinicus* (Indian Mint), *Brassica Juncea* (Oriental mustard), and many more.

The next stage is data preparation. The first sub-phase of data preparation is image normalization. This process is done by multiplying each pixel value by 1./255. The second data preparation stage is image augmentation. This stage is carried out by applying several image augmentation techniques to obtain additional synthetic data [32], [33]. The augmentations performed are `horizontal_flip`, `vertical_flip`, `width_shift`, `height_shift`, `rotation`, `fill_mode = 'reflect'`, `zoom`, `brightness_range = [0.5, 1.5]`, `featurewise_std_normalization = True`, `shear` and `featurewise_center` [34]–[37]. There are two stages of feature extraction carried out, as shown in the Fig. 4.

In data preparation phase, the dataset folder is named according to the scientific name of the species. The dataset is broken down for data training, validation, and testing. The entire dataset has been segmented to free from the background. The distribution of training and validation data is 80% for training data and 20% for validation data, with 1500 images of thirty species. The 80/20 dataset composition is based on previous research [38]–[45]. Data testing uses new data outside the dataset for training and validation. The testing data consisted of a total of 43 images of 30 species. Each species class consists of 1-3 images.

The third phase is feature extraction. This study conducted experiments to compare three pre-trained models for feature extraction on medicinal plant image datasets, namely VGG16, VGG19, and MobileNetV2.

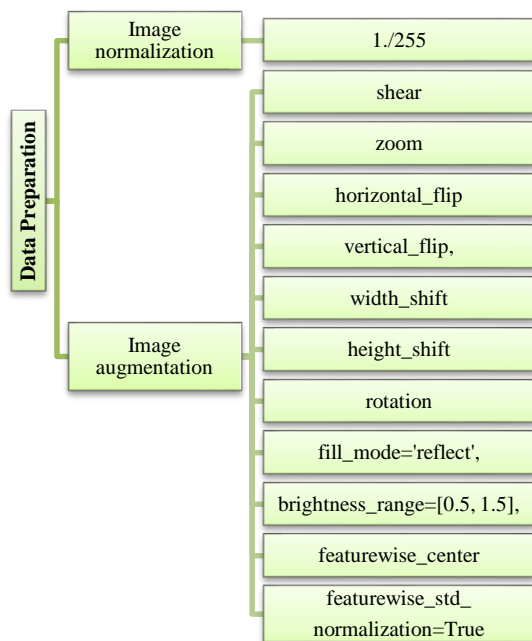


Fig. 4. Feature Extraction Methods.

The fourth phase is classification. The data that has been extracted features a pre-trained model, then training and model validation is carried out using training and validation data. In order to obtain better accuracy results, experiments were also conducted to compare the model with fine-tuning and without fine-tuning. The VGG16, VGG19, and MobileNetV2 model architectures study leaf shape information to differentiate different plant species. The input leaf size and color channel used are adjusted for the VGG16, VGG19, and MobileNetV2 models.

The fifth phase is testing that is done using new data outside the dataset. The testing data obtained by searching through Google using the keyword species name of plant. The data selected on Google is only focused the leaves object, if there is a background in the image, only the leaves are taken (cropped). The testing data consisted of a total of 43 images of 30 species. Each species class consists of 1-3 images. The following Fig. 5 is an example of testing data.

The final phase is evaluation. The evaluation is carried out by comparing the experimental results to find the best suitable model in the dataset. We evaluated the VGG16, VGG19, and MobileNetV2 models for leaf identification on medicinal plant leaves by conducting experiments on the collected datasets. The evaluation method used is the accuracy method. The evaluation was carried out in two stages: evaluation at the training stage and evaluation at the validation stage.



Fig. 5. Example of Data Testing.

#### IV. RESULT AND DISCUSSION

This study used the VGG16, VGG19, and MobileNetV2 classification models for leaf identification on medicinal plant leaves. Two different processes are carried out on the same classification model and dataset, namely with and without fine-tuning implementation. From the implementation results, we want to know how the effect of implementation on the accuracy results during the training, validation, and testing processes using the same dataset and classification model.

The first stage is the training process. The training process is carried out by conducting experiments on 80% of the data from the dataset as a whole. The dataset consists of 1500 images (for 30 classes), meaning that there are about 1200 data used in this experiment. The experimental results can be seen in the Table I.

TABLE I. ACCURACY RESULTS ON TRAINING DATA

Model	Training Experiment	
	Without Fine Tuning	With Fine Tuning
VGG16	99.24%	95.72%
VGG19	93.85%	95.51%
MobileNetV2	98.89%	98.41%

Based on the data in Table I, the model without fine-tuning obtains an accuracy of 99.24% for the VGG16 model, 93.85% for the VGG19 model, and 98.89% for the MobileNetV2 model. In contrast, the models with fine-tuning get 95.72% accuracy for the VGG16 model, 95.51% for the VGG19 model, and 98.41% for the MobileNetV2 model. Moreover, the difference Value of training experiment of model implementation using the fine-tuning and without fine-tuning is depicted in the Fig. 6.

Based on the data in Table II, the model without fine-tuning obtains an accuracy of 95.17% for the VGG16 model, 89.49% for the VGG19 model, and 87.50% for the MobileNetV2 model. While the models with fine-tuning get 93.75% accuracy for the VGG16 model, 92.90% for the VGG19 model, and 96.02% for the MobileNetV2 model. Moreover, difference value of validation experiment of model implementation by using fine tuning and without tuning is depicted Fig. 7.

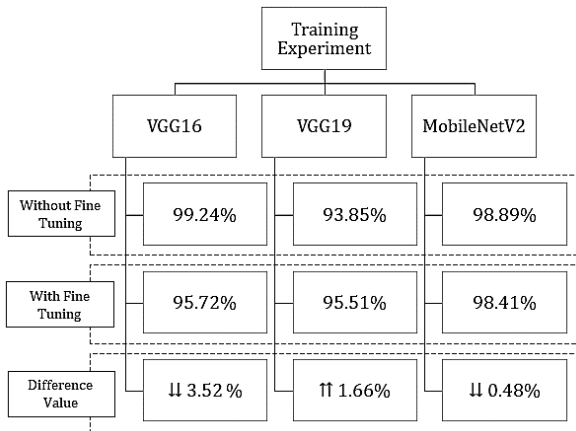


Fig. 6. Difference Value of Training Experiment.

The second stage is the validation process. The validation process is carried out by conducting experiments on 20% of the data from the dataset as a whole. The dataset consists of 1500 images (for 30 classes), meaning that about 300 pieces of data are used in this experiment. The experimental results can be seen in the Table II.

TABLE II. DATA ACCURACY RESULTS ON DATA VALIDATION

Model	Validation Experiment	
	Without Fine Tuning	With Fine Tuning
VGG16	95.17%	93.75%
VGG19	89.49%	92.90%
MobileNetV2	87.50%	96.02%

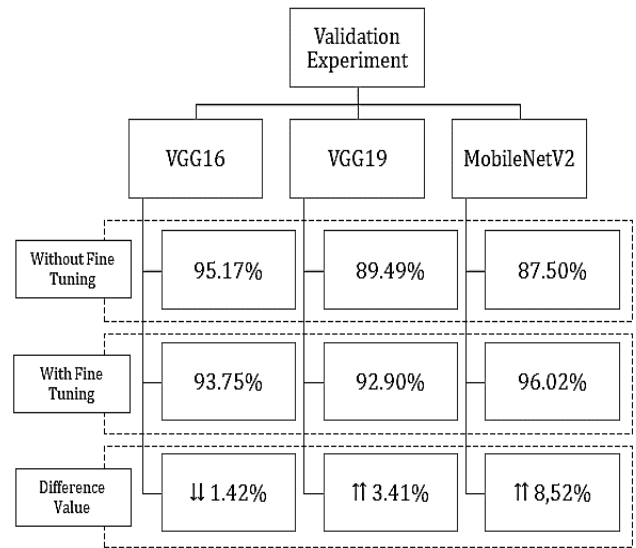


Fig. 7. Difference Value of Validation Experiment.

Based on the Fig. 7, MobileNetV2 and VGG19 with fine-tuning can significantly increase the validation accuracy, while in VGG16 with fine-tuning, it reduces validation accuracy. The third stage is the testing process. Testing data is new data that the model has never processed. After the model is obtained, then testing is carried out on the model using data testing and the result can be seen in Table III.

TABLE III. ACCURACY RESULTS ON TESTING DATA

Model	Testing Experiment	
	Without Fine Tuning	With Fine Tuning
VGG16	22.73%	36.36%
VGG19	15.91%	31.82%
MobileNetV2	43.18%	81.82%

From the experiment results, the models with fine-tuning both VGG16, VGG19, and MobileNetV2 experienced an increase in testing accuracy compared to the model before fine-tuning and the result can be seen in Table IV.

MobileNetV2 obtained the best model by fine-tuning with 96.02% validation accuracy, 81.82% testing accuracy, precision, recall, and f1-score values 0.73, 0.82, and 0.76. The following Fig. 8 is a classification report and confusion matrix obtained by the MobileNetV2 model by fine-tuning.

Overall, VGG16 obtained the highest accuracy compared to other models during the experiment of training without fine tuning (99.24%) and validation without fine tuning (95.17%). While in other experiments the MobileNetV2 model is superior to other models, as shown in the Table V.

Based on the experiment result, we recommended MobileNetV2 model to identify medicinal plant leaves according to the dataset that has been collected. MobileNetV2 was chosen for further research because it got the best accuracy and shorter computation time in recognizing the image of medicinal plant leaves [46]–[48].

TABLE IV. DETAIL OF RESULTS ON TESTING DATA

Class	Precision	Recall	F1-Score
Alpinia Galanga	1.00	1.00	1.00
Amaranthus Viridis	0.67	1.00	0.80
Artocarpus Heterophyllus	1.00	1.00	1.00
Azadirachta Indica	1.00	1.00	1.00
Basella Alba	1.00	1.00	1.00
Brassica Juncea	1.00	1.00	1.00
Carissa Carandas	0.00	0.00	0.00
Citrus Limon	1.00	1.00	1.00
Ficus Auriculata	1.00	1.00	1.00
Ficus Religiosa	1.00	1.00	1.00
Hibiscus Rosa-Sinesis	1.00	1.00	1.00
Jasminum	0.60	1.00	0.75
Mangifera Indica	0.67	1.00	0.80
Mentha	1.00	1.00	1.00
Moringa Oleifera	0.00	0.00	0.00
Muntingia Calabura	0.33	1.00	0.50
Muraya Koenigii	0.00	0.00	0.00
Nerium Oleander	0.00	0.00	0.00
Nyctanthes Arbor-tristis	1.00	1.00	1.00
Ocimum Tenuiflorum	0.00	0.00	0.00
Piper Betle	1.00	1.00	1.00
Plectranthus Amboinicus	0.00	0.00	0.00
Pongamis Pinnata	1.00	1.00	1.00
Psidium Guajava	1.00	1.00	1.00
Punica Granatum	1.00	1.00	1.00
Santalum Album	1.00	1.00	1.00
Syzygium Cumini	0.50	1.00	0.67
Syzygium Jambos	0.00	0.00	0.00
Tabernaemontana Divaricata	1.00	1.00	1.00
Trigonella Foenum-graecum	0.00	0.00	0.00
<b>Macro avg</b>	<b>0.66</b>	<b>0.66</b>	<b>0.66</b>
<b>Weighted avg</b>	<b>0.73</b>	<b>0.73</b>	<b>0.73</b>

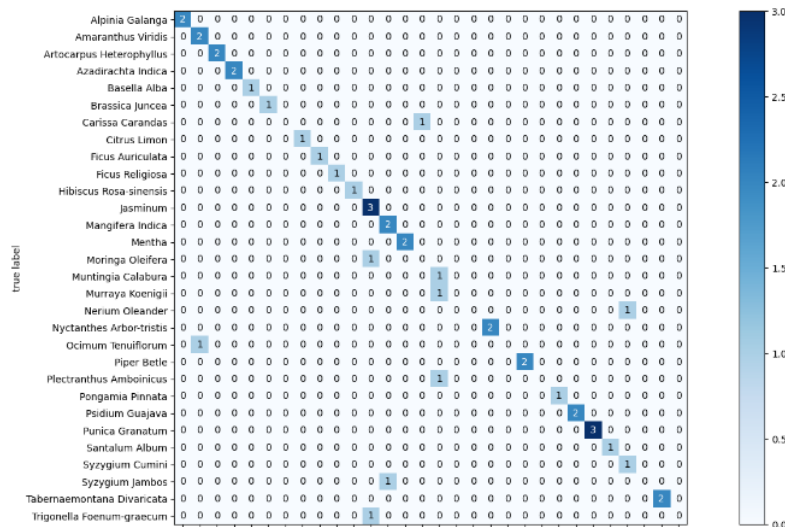


Fig. 8. Confusion Matrix.

TABLE V. OVERALL RESULT OF EXPERIMENT

Phase	Method	VGG16	VGG19	MobileNet
TR	WFT	99.24%	93.85%	98.89%
TR	FT	95.72%	95.51%	98.41%
VA	WFT	95.17%	89.49%	87.50%
VA	FT	93.75%	92.90%	96.02%
TE	WFT	22.73%	15.91%	43.18%
TE	FT	36.36%	31.82%	81.82%

\*WFT = without fine tuning, FT = with fine tuning, TR=training, VA=validation, TE=testing

## V. CONCLUSION

MobileNetV2 obtained the best validation accuracy with fine-tuning with a validation accuracy of 96.02%. MobileNetV2 also obtained the best testing accuracy with fine-tuning of 81.82%. In other models, the testing accuracy obtained is far below MobileNetV2. This condition is likely to happen because the dataset for training and validation used is less diverse and general to build a good model, so the resulting model overfits the dataset. In this case, MobileNetV2 with fine-tuning is quite able to overcome the weaknesses of the dataset used so that when new testing data is given, the accuracy results obtained are quite good. In addition, based on the experiment results, fine-tuning the model can improve the accuracy of the validation and testing produced.

The limitation of this study is that it ignores the background problem of the leaf image. Further research will be carried out using a dataset with a complex background by adding a segmentation method before being processed by the MobileNetV2 model.

## ACKNOWLEDGMENT

The authors would like to thank Universitas Sriwijaya and Universitas Mercu Buana that have been supported this research.

## REFERENCES

- [1] X.-F. Wang, D.-S. Huang, J.-X. Du, H. Xu, and L. Heutte, "Classification of plant leaf images with complicated background," *Appl. Math. Comput.*, vol. 205, no. 2, pp. 916–926, Nov. 2008.
- [2] K. Yang, W. Zhong, and F. Li, "Leaf Segmentation and Classification with a Complicated Background Using Deep Learning," *Agronomy*, vol. 10, no. 11, p. 1721, Nov. 2020.
- [3] I. Ranggadara, Y. S. Sari, S. Dwiasnati, and I. Prihandi, "A Review of Implementation and Obstacles in Predictive Machine Learning Model at Educational Institutions," in *Journal of Physics: Conference Series*, 2020, vol. 1477, p. 32019.
- [4] Y. Jumaryadi, D. Firdaus, B. Priambodo, and Z. P. Putra, "Determining the Best Graduation Using Fuzzy AHP," in *2020 2nd International Conference on Broadband Communications, Wireless Sensors and Powering (BCWSP)*, 2020, pp. 59–63.
- [5] E. M. Jawad, H. G. Daway, and H. J. Mohamad, "Quantum-dot Cellular Automata Based Lossless CFA Image Compression Using Improved and Extended Golomb-rice Entropy Coder," *Int. J. Intell. Eng. Syst.*, vol. 15, no. 2, pp. 12–25, Apr. 2022.
- [6] V. Ayumi, "Pose-based human action recognition with Extreme Gradient Boosting," in *Proceedings - 14th IEEE Student Conference on Research and Development (SCORED)*, 2016, pp. 1–5.
- [7] V. Ayumi, "Mobile Application for Monitoring of Addition of Drugs to Infusion Fluids," *Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol.*, pp. 48–56, Nov. 2019.

- [8] V. Ayumi, "Pose-based human action recognition with Extreme Gradient Boosting," in *Proceedings - 14th IEEE Student Conference on Research and Development: Advancing Technology for Humanity, SCORED*, 2017.
- [9] V. Ayumi, "Application of Machine Learning for SARS-CoV-2 Outbreak," *Int. J. Sci. Res. Sci. Eng. Technol.*, vol. 7, no. 5, pp. 241–248, Oct. 2021.
- [10] V. Ayumi, "Performance Evaluation of Support Vector Machine Algorithm for Human Gesture Recognition," *Int. J. Sci. Res. Sci. Eng. Technol.*, vol. 7, no. 6, pp. 204–210, 2020.
- [11] J. Arun Pandian, G. Geetharamani, and B. Annette, "Data Augmentation on Plant Leaf Disease Image Dataset Using Image Manipulation and Deep Learning Techniques," in *2019 IEEE 9th International Conference on Advanced Computing (IACC)*, 2019, pp. 199–204.
- [12] J. S. Cope, D. Corney, J. Y. Clark, P. Remagnino, and P. Wilkin, "Plant species identification using digital morphometrics: A review," *Expert Syst. Appl.*, vol. 39, no. 8, pp. 7562–7573, Jun. 2012.
- [13] Y. A. Putri, E. C. Djamal, and R. Ilyas, "Identification of Medicinal Plant Leaves Using Convolutional Neural Network," *J. Phys. Conf. Ser.*, vol. 1845, no. 1, p. 012026, Mar. 2021.
- [14] M. A. F. Azlah, L. S. Chua, F. R. Rahmad, F. I. Abdullah, and S. R. Wan Alwi, "Review on Techniques for Plant Leaf Classification and Recognition," *Computers*, vol. 8, no. 4, p. 77, Oct. 2019.
- [15] F. Ghouse, K. Paranjothi, and R. Vaithianathan, "Dysgraphia Classification based on the Non-Discrimination Regularization in Rotational Region Convolutional Neural Network," *Int. J. Intell. Eng. Syst.*, vol. 15, no. 1, Feb. 2022.
- [16] J. Lu, L. Tan, and H. Jiang, "Review on Convolutional Neural Network (CNN) Applied to Plant Leaf Disease Classification," *Agriculture*, vol. 11, no. 8, p. 707, Jul. 2021.
- [17] E. Korot et al., "Code-free deep learning for multi-modality medical image classification," *Nat. Mach. Intell.*, vol. 3, no. 4, pp. 288–298, Apr. 2021.
- [18] V. Ayumi, E. Ermatita, A. Abdiansah, H. Noprisson, M. Purba, and M. Utami, "A Study on Medicinal Plant Leaf Recognition Using Artificial Intelligence," in *2021 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS, 2021)*, pp. 40–45.
- [19] E. Hidayat, Lukman, H. Noprisson, D. I. Sensuse, Y. G. Sucahyo, and E. D. Putra, "Development of mobile application for documenting traditional knowledge in Indonesia: A Case Study of Traditional Knowledge in Using Medicinal Plant," in *Proceedings - 14th IEEE Student Conference on Research and Development: Advancing Technology for Humanity, SCORED 2016*, 2017.
- [20] H. Noprisson, D. I. Sensuse, Y. G. Sucahyo, and Lukman, "Metadata Development for Ethnophytomedicine Resources Using Metadata Analysis Approach," in *The 2016 8th International Conference on Information Technology and Electrical Engineering (ICITEE 2016)*, 2016.
- [21] D. I. Sensuse, H. Noprisson, Y. G. Sucahyo, and L. Lukman, "Metadata Schema for Traditional Knowledge," *Int. J. Adv. Sci. Eng. Inf. Technol.*, vol. 7, no. 6, 2017.
- [22] Y. G. Naresh and H. S. Nagendraswamy, "A novel fuzzy LBP based symbolic representation technique for classification of medicinal plants," in *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*, 2015, pp. 524–528.
- [23] G. Mukherjee, A. Chatterjee, and B. Tudu, "Study on the potential of combined GLCM features towards medicinal plant classification," in *2016 2nd International Conference on Control, Instrumentation, Energy & Communication (CIEC)*, 2016, pp. 98–102.
- [24] D. Venkataraman and N. Mangayarkarasi, "Support vector machine based classification of medicinal plants using leaf features," in *2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, 2017, pp. 793–798.
- [25] L. Gao and X. Lin, "A method for accurately segmenting images of medicinal plant leaves with complex backgrounds," *Comput. Electron. Agric.*, vol. 155, pp. 426–445, 2018.
- [26] C. Sivaranjani, L. Kalinathan, R. Amutha, R. S. Kathavarayan, and K. J. Jegadish Kumar, "Real-Time Identification of Medicinal Plants using

- Machine Learning Techniques,” in 2019 International Conference on Computational Intelligence in Data Science (ICCIDS), 2019, pp. 1–4.
- [27] D. Pechebovicz et al., “Plants recognition using embedded Convolutional Neural Networks on Mobile devices,” in 2020 IEEE International Conference on Industrial Technology (ICIT), 2020, pp. 674–679.
- [28] M. R. Bhuiyan, M. Abdullahil-Oaphy, R. S. Khanam, and M. S. Islam, “MediNET: A Deep Learning Approach to Recognize Bangladeshi Ordinary Medicinal Plants Using CNN,” in *Soft Computing Techniques and Applications*, Springer, 2021, pp. 371–380.
- [29] S. Bhattacharya, A. Mukherjee, and S. Phadikar, “A deep learning approach for the classification of rice leaf diseases,” in *Intelligence enabled research*, Springer, 2020, pp. 61–69.
- [30] S. Roopashree and J. Anitha, “Medicinal leaf dataset,” Mendeley Data, vol. 1, 2020.
- [31] R. Shailendra et al., “An IoT and Machine Learning Based Intelligent System for the Classification of Therapeutic Plants,” *Neural Process. Lett.*, pp. 1–29, 2022.
- [32] M. Defriani and I. Jaelani, “Recognition of Regional Traditional House in Indonesia Using Convolutional Neural Network (CNN) Method,” *J. Comput. Networks, Archit. High Perform. Comput.*, vol. 4, no. 2, pp. 104–115, 2022.
- [33] R. M. Rao and M. K. Arora, “Overview of image processing,” in *Advanced image processing techniques for remotely sensed hyperspectral data*, Springer, 2004, pp. 51–85.
- [34] S. Mishra, D. Koner, L. Jena, and P. Ranjan, “Leaves shape categorization using convolution neural network model,” in *Intelligent and cloud computing*, Springer, 2021, pp. 375–383.
- [35] V. Maslej-Krešňáková, K. El Boucheffy, and P. Butka, “Morphological classification of compact and extended radio galaxies using convolutional neural networks and data augmentation techniques,” *Mon. Not. R. Astron. Soc.*, vol. 505, no. 1, pp. 1464–1475, 2021.
- [36] I. Bhakta, S. Phadikar, and K. Majumder, “Thermal Image Augmentation with Generative Adversarial Network for Agricultural Disease Prediction,” in *International Conference on Computational Intelligence in Pattern Recognition*, 2022, pp. 345–354.
- [37] R. Karothia and M. K. Chattopadhyay MIEEE, “Vigorous Deep Learning Models for Identifying Tomato Leaf Diseases,” in *Proceedings of International Conference on Data Science and Applications*, 2022, pp. 131–152.
- [38] S. Widiyanto, R. Fitrianto, and D. T. Wardani, “Implementation of Convolutional Neural Network Method for Classification of Diseases in Tomato Leaves,” in 2019 Fourth International Conference on Informatics and Computing (ICIC), 2019, pp. 1–5.
- [39] P. Bharali, C. Bhuyan, and A. Boruah, “Plant disease detection by leaf image classification using convolutional neural network,” in *International conference on information, communication and computing technology*, 2019, pp. 194–205.
- [40] M. M. Ghazi, B. Yanikoglu, and E. Aptoula, “Plant identification using deep neural networks via optimization of transfer learning parameters,” *Neurocomputing*, vol. 235, pp. 228–235, 2017.
- [41] M. S. H. Kalathingal, S. Basak, and J. Mitra, “Artificial neural network modeling and genetic algorithm optimization of process parameters in fluidized bed drying of green tea leaves,” *J. Food Process Eng.*, vol. 43, no. 1, p. e13128, 2020.
- [42] M. Syarief and W. Setiawan, “Convolutional neural network for maize leaf disease image classification,” *Telkomnika (Telecommunication Comput. Electron. Control.*, vol. 18, no. 3, pp. 1376–1381, 2020.
- [43] R. Thangaraj, S. Anandamurugan, and V. K. Kaliappan, “Automated tomato leaf disease classification using transfer learning-based deep convolution neural network,” *J. Plant Dis. Prot.*, vol. 128, no. 1, pp. 73–86, 2021.
- [44] M. Ji, L. Zhang, and Q. Wu, “Automatic grape leaf diseases identification via UnitedModel based on multiple convolutional neural networks,” *Inf. Process. Agric.*, vol. 7, no. 3, pp. 418–426, 2020.
- [45] S. A. Wagle, R. Harikrishnan, S. H. M. Ali, and M. Faseehuddin, “Classification of plant leaves using new compact convolutional neural network models,” *Plants*, vol. 11, no. 1, p. 24, 2021.
- [46] H.-H. Yen et al., “Performance comparison of the deep learning and the human endoscopist for bleeding peptic ulcer disease,” *J. Med. Biol. Eng.*, vol. 41, no. 4, pp. 504–513, 2021.
- [47] C. H. Karadal, M. C. Kaya, T. Tuncer, S. Dogan, and U. R. Acharya, “Automated classification of remote sensing images using multileveled MobileNetV2 and DWT techniques,” *Expert Syst. Appl.*, vol. 185, p. 115659, 2021.
- [48] P. N. Huu, N. N. Thi, and T. P. Ngoc, “Proposing Posture Recognition System Combining MobilenetV2 and LSTM for Medical Surveillance,” *IEEE Access*, vol. 10, pp. 1839–1849, 2021.

# Effect of Random Splitting and Cross Validation for Indonesian Opinion Mining using Machine Learning Approach

Mariana Purba<sup>1</sup>

Doctor of Engineering  
Universitas Sriwijaya, Indonesia

Handrie Noprisson<sup>4</sup>, Vina Ayumi<sup>5</sup>, Umniy Salamah<sup>7</sup>

Faculty of Computer Science  
Universitas Mercu Buana, Indonesia

Ermatita Ermatita<sup>2</sup>, Abdiansah Abdiansah<sup>3</sup>

Faculty of Computer Science  
Universitas Sriwijaya, Indonesia

Hadiguna Setiawan<sup>6</sup>

Department of Computer Science  
Stikhafi Indonesia

Yadi Yadi<sup>8</sup>

Informatics Department  
Institut Teknologi Pagar Alam, Indonesia

**Abstract**—Opinion mining has been a prominent topic of research in Indonesia, however there are still many unanswered questions. The majority of past research has been on machine learning methods and models. A comparison of the effects of random splitting and cross-validation on processing performance is required. Text data is in Indonesian. The goal of this project is to use a machine learning model to conduct opinion mining on Indonesian text data using a random splitting and cross validation approach. This research consists of five stages: data collection, pre-processing, feature extraction, training & testing, and evaluation. Based on the experimental results, the TF-IDF feature is better than the Count-Vectorizer (CV) for Indonesian text. The best accuracy results are obtained by using TF-IDF as a feature and Support Vector Machine (SVM) as a classifier with cross validation implementation. The best accuracy reaches 81%. From the experimental results, it can also be seen that the implementation of cross validation can improve accuracy compared to the implementation of random splitting.

**Keywords**—Random splitting; cross validation; machine learning; Indonesian text

## I. INTRODUCTION

Opinion mining technology examines and interprets enormous amounts of text data automatically. Opinion mining is the technique of extracting useful information and knowledge from unstructured natural language texts automatically [1]–[5]. Opinion mining is a specific sub-field of text mining that seeks to automatically discover the polarity of opinions (positive, negative, and others) associated with natural language texts [6]–[11].

The language structure of the dataset determines the main challenge in opinion mining. Sentences can be ironic or have several meanings depending on the context. For example, someone can support school policies in the education sector

while also breaking school rules—another challenge in obtaining opinions in determining the difference between subjective and objective texts [12]–[14]. A subjective text is one person's point of view, bias, or opinion. News stories and neutral texts are examples of objective writing that deal with facts and are supposed to be fully unbiased [6], [15], [16].

A machine learning approach can be used for opinion mining. Machine learning refers to methods and systems that can learn from data automatically. The most common machine learning method is supervised learning. It entails creating a prediction model that can inductively learn from a training data collection [17]. The training data is a set of labelled instances, with each example consisting of a pair of input objects (specified in a feature set) and the desired output value, in the case of a classification model, a class label. After the model has been trained, it is ready to be applied to new data [6].

Opinion mining in Indonesia has been a popular topic of study, yet there are still many open challenges. Indonesia is morphologically diverse and ambiguous, with complicated morpho-syntactic rules and many irregular forms and a wide range of dialects with no written standards. Learning a robust general model from Indonesian text might be challenging without suitable processing and handling [18]. Furthermore, compared to English, Indonesian opinion mining has fewer freely available resources in terms of huge net sentiment lexicons and annotated opinion sets. These difficulties have piqued researchers' interest in Indonesian opinion mining [19].

Apart from increasing research on opinion mining on Indonesian text data, there are still some gaps. Most of the previous research has focused on machine learning models and algorithms. There is a need to compare the effect of random splitting and cross-validation on improving performance for processing Indonesian text data [19], [20].



Based on the above background, the purpose of this study is to conduct opinion mining on Indonesian text data using a machine learning model by implementing a random splitting and cross validation approach.

## II. RELATED WORK

The previous research of opinion mining on Indonesian text data using a machine learning model has been done by several scholars. Research by Fachrina & Widyanoro (2017) compares Support Vector Machine and Naïve Bayes Classifier to process 2960 Indonesian text data [21]. Research by Suciati & Budi (2019) compared the performance of Random Forest (RF), Multinomial Naïve Bayes (NB), Logistic Regression (LR), Decision Tree (DT), and Extra Tree classifier (ET) using ten folds cross-validation to process Indonesian text data.

The algorithm that achieved the highest score was obtained by Logistic Regression (LR) and Decision Tree (DT) [22]. Research by Miranda et al. (2019) used Bayes classification to process Indonesian text data. This study obtained an accuracy of 74.94% [23]. Research by Wisnu et al. (2020) uses the Naïve Bayes classifier and K-Nearest Neighbor or KNN to process Indonesian text data. This study obtained an accuracy of KNN (91.00%) and Naïve Bayes (83.50%) [24]. Research by Buntoro et al. (2021) uses the Naïve Bayes Classifier (NBC) and a Support Vector Machine (SVM) to process Indonesian text data. This study obtained an accuracy of SVM (79.02%) and NBC (44.94%) [25].

Most of the previous research has focused on machine learning models and algorithms. There is a need to compare the effect of random splitting and cross-validation on improving performance for processing Indonesian text data. This research is proposed to fill the gap by implementing random splitting and cross-validation for improving performance for processing Indonesian text data.

## III. RESEARCH METHODOLOGY

This study uses a public dataset to determine negative and positive comments on the Indonesian feedback dataset. The stages of the research can be seen in the following Fig. 1.

The first stage is the data collection stage. The dataset used for training and testing the model is sentiment data on Twitter obtained publicly provided by Sulistya in 2021 at Kaggle [26]. The dataset is a collection of feedback data in Indonesian by Twitter users on Covid-19. The dataset consists of 1000 records with 500 records each for the positive class and 500 tweets for the negative class. The following is an example of a dataset. The second stage is the preprocessing stage. This stage consists of six stages: data cleansing, case folding, tokenizing, stopword, normalization, and stemming. Details of these stages can be seen in the Fig. 2:

Based on Fig. 2 above, at the data cleansing stage, a cleaning process is carried out for words that are not needed in order to reduce the computational burden, such as text containing HTML, links, and scripts. In addition, this stage also removes punctuation marks such as periods (.), commas (,) and other punctuation marks. In this pre-processing process, the case folding method is also applied, namely the process of converting words into lower-case letters. The third stage is

tokenization. This method is implemented to transform the text's words into several sequences truncated by spaces or specific characters [23], [27], [28].

The stop word removal method is a method of deleting a word with a unique word from text data such as conjunctions and possessive words. In addition, types of words that are less meaningful will be removed, such as words: I, and, or by using this method. Stop word removal is meant to reduce the burden on system performance because the words taken are considered essential [29]–[31]. The last stage in the pre-processing process is the stemming stage. The method at this stage is done by transforming the words in the text to become essential words.

The third research stage is to perform feature extraction. We compare two text features at this stage, namely Count Vectorizer (CV) and TF-IDF. The Term Frequency-Inverse Document Frequency method, abbreviated as TF-IDF, is the most commonly used word weight calculation method in opinion mining. The method is known for its efficient computation time, ease of implementation, and good results or accuracy values. The method works by calculating the value of Term Frequency (TF) and Inverse Document Frequency (IDF) for each token (word) in the document in the corpus or dataset [32]–[34].

The fourth stage is the Training and Testing Model. Training and testing are done by comparing four classifiers, namely Random Forest (RF), Logistic Regression (LR), Naïve Bayes (NB), and Support Vector Machine (SVM)

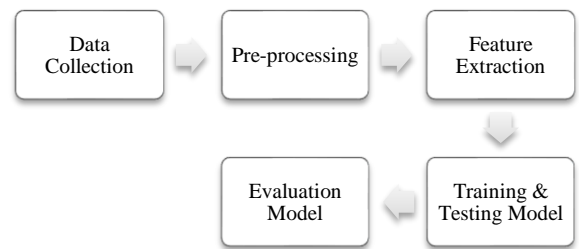


Fig. 1. Main Research Methods.

<b>Data Cleansing</b>	<ul style="list-style-type: none"> <li>• Hashtag (#) and mention (@), URL, punctuation, emoticon</li> </ul>
<b>a. Case Folding</b>	<ul style="list-style-type: none"> <li>• Convert text into lowercase one</li> </ul>
<b>a. Tokenizing</b>	<ul style="list-style-type: none"> <li>• White space and punctuation as token delimiters</li> </ul>
<b>Stopword Removal</b>	<ul style="list-style-type: none"> <li>• Stopword removal based on Indonesian dictionary</li> </ul>
<b>Normalization</b>	<ul style="list-style-type: none"> <li>• Convert slang word to formal word</li> </ul>
<b>Stemming</b>	<ul style="list-style-type: none"> <li>• Convert prepositional words to base words</li> </ul>

Fig. 2. Pre-Processing Methods.

The Random Forest (RF) method is the development of the Classification and Regression Tree (CART) algorithm. This method applies bootstrap aggregating and random feature selection. This method can be used for classification that works by building a classification tree. Increasing the accuracy of the RF method is done by generating a child node on each node (the node above it) with random selection. Then, the classification results from each tree are accumulated and selected based on the classification results that appear the most [35]–[37]. The RF method has three main parts: the root node, internal node, and leaf node. The root node is the node at the very top, commonly referred to as the root of the decision tree. The internal node is the branch's node with one to two inputs. Finally, the leaf node or terminal node is the end node that has one input and no output. The calculation on the decision tree begins by calculating the entropy value as a determinant of the level of attribute impurity and the value of information gain [35]–[37]. The Logistic Regression (LR) method is used to express the relationship between categorical response variables (in the form of polychotomous or dichotomous) with either continuous or categorical predictor variables. Logistic regression aims to classify each event or observation into positive and negative classes [38] [39].

The Naïve Bayes (NB) method is a method that can be used in opinion mining. This method applies Bayes' theorem theory in classification based on attribute values that are conditionally independent if given an output value. In short, Bayes' theorem is a fundamental statistical approach to pattern recognition [40], [41]. The advantage of using the NB method is that the value or amount of training data required in data processing can be on a small scale and can still be used to determine parameter estimates in the data classification process [42].

The Support Vector Machine (SVM) method is a method with the concept of statistical learning theory, which has given better performance results than other classification methods in several research cases. This method does not study all the training data in the training process, but only a selected number of data is used to build a model in the classification process. The SVM method does not store all training data but only stores a small portion of training data for further processing. This has become an advantage in choosing the SVM method because not all training data is involved in each training iteration [43]. The SVM method works by maximizing the decision limit (hyperplane) or finding the best decision limit (hyperplane) as a separator of the two data classes depicted in the Fig. 3:

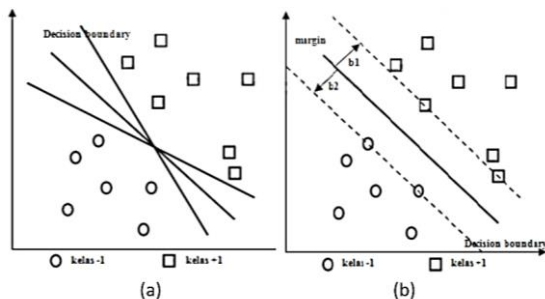


Fig. 3. Hyperplane [44].

In the picture above (a), there is a choice of possible decision limits (hyperplane) for the data set; then, in the picture above (b), there is a decision limit (hyperplane) with the maximum margin. The margin is the distance between the decision boundary (hyperplane) and the closest data from each class. This closest data is known as the support vector. The hyperplane component with the maximum margin will better generalize the classification process.

For example, in Fig 3 (b), the solid line component shows the best decision boundary (hyperplane) with a location in the middle of the two classes, while the dotted line component that passes through the circle and square data is a support vector. The central concept of training on the SVM method is finding the hyperplane's location [44], [45]. Experiments were carried out using the results of random splitting and cross-validation. The third and fourth stages are illustrated in the Fig. 4:

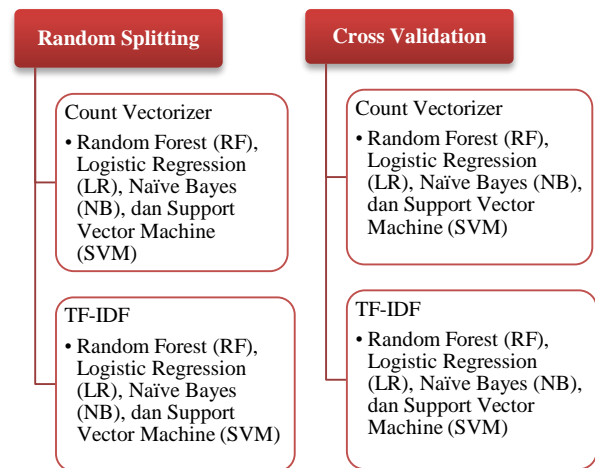


Fig. 4. Experiment Scenario.

The next stage is an evaluation to compare the best accuracy, precision, recall, f1-measure obtained. The performance evaluation measurement model is an approach that aims to measure the performance or performance of a system. This approach is widely used in the case of training or data training. Several formulas or equations in the performance evaluation measure are usually applied separately or in combination to get a better performance analysis perspective. Some of the calculations contained in the performance evaluation are as follows [46]. The precision method calculates the level of accuracy or accuracy of the results between user testing and system answers.

$$pre = \frac{TP}{TP+FP} \tag{1}$$

The recall is a measurement of the accuracy or accuracy of the same information with information that has existed before.

$$rec = \frac{TP}{TP+FN} \tag{2}$$

Accuration is a comparative calculation between the information the system answers correctly with the comprehensive information.

$$acc = \frac{TP+TN}{TP+TN+FP+FN} \tag{3}$$

IV. RESULT AND DISCUSSION

This study aims to determine negative and positive comments on the Indonesian feedback dataset publicly provided by Sulistya (2021) on the Kaggle web using a machine learning approach. This study also compares the performance of several machine learning methods to find out which method has the best performance.

The second stage carried out in the experiment is the pre-processing stage. The first stage of pre-processing is data cleansing. The cleansing of the dataset is: removing hashtags (#) and mentions (@), deleting URLs, deleting punctuation and deleting emoticons. The example result of data cleansing can be seen in the Table I:

TABLE I. RESULT OF DATA CLEANSING

Process	Data #1	Data #2
Data source	"Indonesia: APBN Sekarat, Covid-19 Meningkat,	#BREAKING:Pemerintah mengonfirmasi kasus posi
Text_remove_hastag_and mentions	"Indonesia:APBN Sekarat, Covid-19 Meningkat,	: Pemerintah mengonfirmasi kasus positif Covid
Text_remove_url	"Indonesia:APBN Sekarat, Covid-19 Meningkat,	: Pemerintah mengonfirmasi kasus positif Covid
Text_remove_punc	Indonesia:APBN Sekarat, Covid-19 Meningkat Rakyat	Pemerintah mengonfirmasi kasus positif Covid1
Text_remove_emojis	Indonesia:APBN Sekarat, Covid-19 Meningkat Rakyat	Pemerintah mengonfirmasi kasus positif Covid1
Text_remove_emoticons	Indonesia:APBN Sekarat, Covid-19 Meningkat Rakyat	Pemerintah mengonfirmasi kasus positif Covid1

The second preprocessing stage is case folding. This stage is done by converting text into lowercase ones. For example, "Pemerintah mengkonfirmasi kasus positif COVID19...", will be converted as "pemerintah mengkonfirmasi kasus covid19". The example results of case folding for our dataset can be seen in the Fig. 5:

```

0 bang gimna pemerintah mau peduli rrc urus abk ...
1 erinx tidak percaya data covid19 dari pemerin...
2 indonesia apbn sekarat covid19 meningkat rakya...
3 \n\nuntuk mengurangi penyebaran virus covid19\...
4 hingga Kamis 752020 pk 1200 Wibdata pemerint...
5 pemerintah mengonfirmasi kasus positif covid1...
```

Fig. 5. Example Result of Case Folding.

The next stage is tokenizing. This stage separates the text with white space and punctuation as token delimiters. For example, "pemerintah mengkonfirmasi kasus covid19", will be converted as "[pemerintah, mengkonformasi, kasus, positif, covid19]". The example result of tokenizing for our dataset can be seen in the Fig. 6:

```

0 [bang, gimna, pemerintah, mau, peduli, rrc, ur...
1 [erinx, tidak, percaya, data, covid, dari, pem...
2 [indonesia, apbn, sekarat, covid, meningkat, r...
3 [untuk, mengurangi, penyebaran, virus, covid, ...
4 [hingga, Kamis, pk, Wibdata, pemerintah, mempe...
5 [pemerintah, mengonfirmasi, kasus, positif, co...
```

Fig. 6. Example Results of Tokenizing.

The next step is stop-word removal. At this stage, the words included in the stop-word will be deleted. Stop-word deletion is done by matching the dataset with the Indonesian stop-word dictionary. For example, "[untuk, mengurangi, penyebaran virus, covid]", will be converted "[mengurangi, penyebaran virus, covid]". The word "untuk" stop-word so as it will be deleted. The example result of stop-word removal can be seen in the Fig. 7:

case_folding_tweets	tweet_tokens	tweet_stopword_removal
bang gimna pemerintah mau peduli rrc urus abk ...	[bang, gimna, pemerintah, mau, peduli, rrc, ur...	[bang, gimna, pemerintah, peduli, rrc, urus, a...
erinx tidak percaya data covid dari pemerintah...	[erinx, tidak, percaya, data, covid, dari, pem...	[erinx, percaya, data, covid, pemerintah, perc...
indonesia apbn sekarat covid meningkat rakyat ...	[indonesia, apbn, sekarat, covid, meningkat, r...	[indonesia, apbn, sekarat, covid, meningkat, r...
untuk mengurangi penyebaran virus covid menduk...	[untuk, mengurangi, penyebaran, virus, covid, ...	[mengurangi, penyebaran, virus, covid, menduku...

Fig. 7. Example Results of Stop-Removal.

The fifth preprocessing stage is normalization. Normalization changes non-standard words (slang words) and acronyms into familiar words by matching the dataset with the Indonesian normalization dictionary. The results of some normalization of words in Indonesian are shown in the Table II and Fig. 8:

TABLE II. RESULT OF NORMALIZATION

No.	Real Data	Normalization
1	&	Dan
2	l pun	Satupun
3	7an	Tujuan
4	@	Di
5	Jkt	Jakarta
6	Nasihat	Nasehat
7	SE	Surat edaran
8	Ababil	Abglabil
9	Abis	Habis
10	Ad	Ada

tweet_tokens	tweet_stopword_removal	tweet_normalized
[bang, gimna, pemerintah, mau, peduli, rrc, ur...	[bang, gimna, pemerintah, peduli, rrc, urus, a...	[bang, gimana, pemerintah, peduli, rrc, urus, ...
[erinx, tidak, percaya, data, covid, dari, pem...	[erinx, percaya, data, covid, pemerintah, perc...	[erinx, percaya, data, covid, pemerintah, perc...
[indonesia, apbn, sekarat, covid, meningkat, r...	[indonesia, apbn, sekarat, covid, meningkat, r...	[indonesia, apbn, sekarat, covid, meningkat, r...

Fig. 8. Example Results of Normalization.

The sixth preprocessing stage is stemming. At this stage, the affixed words are transformed into basic words using the literary method. The method at this stage is done by transforming the words in the text to become essential words. At this stage, the transformation of affixed words into basic words is carried out using the Sastrawi library. For example, “[mengurangi, penyebaran virus, covid]”, will be converted “[kurang, sebar virus, covid]. The basic words of “mengurangi” and “menyebarkan” are “kurang” and “sebar. The example result of stemming can be seen in the Fig. 9:

After completing six stages: data cleansing, case folding, tokenization, stopword, normalization, and stemming. The example of every step has been elaborated above. Attribute of Data#1 is text before method is applied and attribute of Data#2 is text after applied method. Moreover, an overview result of preprocessing stages can be seen in the Table III.

The third research stage is to perform feature extraction. At this stage, feature extraction is carried out on the dataset that has been processed in the previous stage. The feature extraction stage aims to obtain features used in model training and testing. We compare two text features at this stage, namely Count Vectorizer (CV) and TF-IDF.

The fourth stage is the training and testing model. Training and testing are done by comparing four classifiers, namely Random Forest (RF), Logistic Regression (LR), Naïve Bayes (NB), and Support Vector Machine (SVM). Experiments were carried out using the approach of random splitting and cross-validation.

tweet_stopword_removal	tweet_normalized	tweet_tokens_stemmed
[bang, gimna, pemerintah, peduli, rrc, urus, a...	[bang, gimana, pemerintah, peduli, rrc, urus, ...	[bang, gimana, perintah, peduli, rrc, urus, ab...
[erinx, percaya, data, covid, pemerintah, perc...	[erinx, percaya, data, covid, pemerintah, perc...	[erinx, percaya, data, covid, perintah, percay...
[mengurangi, penyebaran, virus, covid, menduku...	[mengurangi, penyebaran, virus, covid, menduku...	[kurang, sebar, virus, covid, dukung, anjur, p...

Fig. 9. Example Results of Stemming.

TABLE III. RESULT OF DATA PREPROCESSING

Method	Data #1	Data #2
Cleansing_tweets	Indonesia APBN Sekarat Covid Meningkatkan Rakya...	\nUntuk Mengurangi Penyebaran Virus Covid19\ ...
case_folding_tweets	indonesia apbn sekarat covid meningkat rakyat...	untuk mengurangi penyebaran virus covid menduk...
tweet_tokens	[Indonesia, apbn, sekarat, covid, meningkat r...	[untuk, mengurangi, penyebaran, virus, covid, ...
tweet_stopword_removal	[Indonesia, apbn, sekarat, covid, meningkat, r...	[mengurangi, penyebaran, virus, covid, menduku...
tweet_normalized	[Indonesia, apbn, sekarat, covid, meningkat, r...	[mengurangi, penyebaran, virus, covid, menduku...
tweet_tokens_stemmed	[Indonesia, apbn, sekarat, covid, tingkat, rak...	[kurang, sebar, virus, covid, dukung, anjur, p...

The first experiment is training and evaluating machine learning models using random splitting. This experiment was carried out by randomly dividing the dataset into training and testing data with 80% of the training data and 20% of the testing data, respectively. Furthermore, an evaluation was carried out to compare the best accuracy, precision, recall, f1-measure obtained. The results of this experiment can be seen in the Table IV:

TABLE IV. PERFORMANCE EVALUATION

Process	Accuracy	Precision	Recall	F1
CV&RF	0.76	0.77	0.76	0.76
TF-IDF&RF	0.77	0.77	0.77	0.77
CV&LR	0.75	0.75	0.75	0.75
TF-IDF&LR	0.76	0.76	0.76	0.76
CV&NB	0.75	0.75	0.74	0.75
TF-IDF&NB	0.74	0.75	0.74	0.74
CV & SVM	0.76	0.77	0.76	0.76
TF-IDF&SVM	0.78	0.78	0.78	0.78

The second experiment is training and evaluation of machine learning models using cross-validation. This stage is carried out using 10-fold cross-validation. At this stage, cross-validation is implemented to find the maximum accuracy of the model. After cross-validation, training and model evaluation were carried out to measure the resulting accuracy, precision, recall, f1-measure results. The results of this experiment can be seen in the Table 5:

TABLE V. EXPERIMENT WITH CROSS VALIDATION

Process	Accuracy	Precision	Recall	F1
CV&RF	0.79	0.80	0.79	0.79
TF-IDF&RF	0.78	0.79	0.79	0.79
CV&LR	0.78	0.79	0.79	0.79
TF-IDF&LR	<b>0.81</b>	0.81	0.81	0.81
CV&NB	0.79	0.80	0.79	0.79
TF-IDF&NB	0.79	0.80	0.79	0.79
CV & SVM	0.79	0.79	0.79	0.79
TF-IDF&SVM	<b>0.81</b>	0.81	0.81	0.81

According to the Table above, utilizing TF-IDF as a feature and Support Vector Machine (SVM) as a classifier with cross validation implementation yields the best accuracy results. The highest level of accuracy is 81%. It can also be observed from the experimental results that cross validation can improve accuracy when compared to random splitting. Furthermore, for Indonesian language, the TF-IDF feature outperforms the Count-Vectorizer (CV).

## V. CONCLUSION

Based on the experimental results, the TF-IDF feature is better than the Count-Vectorizer (CV) for Indonesian text. The best accuracy results are obtained by using TF-IDF as a feature and Support Vector Machine (SVM) as a classifier with cross validation implementation. The best accuracy reaches 81%. From the experimental results, it can also be seen that the implementation of cross validation can improve accuracy compared to the implementation of random splitting.

This research has not discussed the problem of negation. In future research, issues that will be investigated further include the implementation of negation handling with the modified



syntactic rule method in the pre-processing process to increase the accuracy of opinion mining.

#### ACKNOWLEDGMENT

The authors would like to thank Universitas Sriwijaya that have supported this research.

#### REFERENCES

- [1] M. Misuraca, G. Scepi, and M. Spano, "Using Opinion Mining as an educational analytic: An integrated strategy for the analysis of students' feedback," *Stud. Educ. Eval.*, vol. 68, p. 100979, 2021.
- [2] R. Annisa and I. Surjandari, "Opinion mining on Mandalika hotel reviews using latent dirichlet allocation," *Procedia Comput. Sci.*, vol. 161, pp. 739–746, 2019.
- [3] E. Sonalitha et al., "Combined Text Mining: Fuzzy Clustering for Opinion Mining on the Traditional Culture Arts Work," *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 8, pp. 294–299, 2020.
- [4] D. Ramayanti et al., "Tuberculosis Ontology Generation and Enrichment Based Text Mining," in 2020 International Conference on Information Technology Systems and Innovation (ICITSI), 2020, pp. 429–434.
- [5] H. Noprisson et al., "Influencing factors of knowledge sharing among students in Indonesia higher educational institutions," in 2016 International Conference on Information Technology Systems and Innovation (ICITSI), 2016, pp. 1–6.
- [6] L. Tavoschi et al., "Twitter as a sentinel tool to monitor public opinion on vaccination: an opinion mining analysis from September 2016 to August 2017 in Italy," *Hum. Vaccin. Immunother.*, vol. 16, no. 5, pp. 1062–1069, 2020.
- [7] P. Rajkumar, "Opinion mining for user experience evaluation model using kernel-naive bayes classification algorithm," *J. Eng. Res.*, 2021.
- [8] C. J. Rameshbhai and J. Paulose, "Opinion mining on newspaper headlines using SVM and NLP," *Int. J. Electr. Comput. Eng.*, vol. 9, no. 3, p. 2152, Jun. 2019.
- [9] H. Noprisson, E. Ermatita, A. Abdiansah, V. Ayumi, M. Purba, and M. Utami, "Hand-Woven Fabric Motif Recognition Methods: A Systematic Literature Review," in 2021 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS), 2021, pp. 90–95.
- [10] V. Ayumi, E. Ermatita, A. Abdiansah, H. Noprisson, M. Purba, and M. Utami, "A Study on Medicinal Plant Leaf Recognition Using Artificial Intelligence," in 2021 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS), 2021, pp. 40–45.
- [11] M. Purba, E. Ermatita, A. Abdiansah, V. Ayumi, H. Noprisson, and A. Ratnasari, "A Systematic Literature Review of Knowledge Sharing Practices in Academic Institutions," in 2021 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS), 2021, pp. 337–342.
- [12] Y. Huang, "Opinion Mining Algorithm Based on the Evaluation of Online Mathematics Course with Python," in 2021 5th International Conference on Electronics, Communication and Aerospace Technology (ICECA), 2021, pp. 1395–1398.
- [13] J. Serrano-Guerrero, F. P. Romero, and J. A. Olivas, "Fuzzy logic applied to opinion mining: a review," *Knowledge-Based Syst.*, vol. 222, p. 107018, 2021.
- [14] S. Zad, M. Heidari, J. H. Jones, and O. Uzuner, "A survey on concept-level sentiment analysis techniques of textual data," in 2021 IEEE World AI IoT Congress (AllIoT), 2021, pp. 285–291.
- [15] A. Easwaran, "Opinion Mining and Emotion Detection in Social Network Data and Student Survey Data in Cloud Environment." The University of North Carolina at Charlotte, 2021.
- [16] S. Sagnika, B. S. P. Mishra, and S. K. Meher, "An attention-based CNN-LSTM model for subjectivity detection in opinion-mining," *Neural Comput. Appl.*, vol. 33, no. 24, pp. 17425–17438, 2021.
- [17] H.-C. Soong, N. B. A. Jalil, R. Kumar Ayyasamy, and R. Akbar, "The Essential of Sentiment Analysis and Opinion Mining in Social Media: Introduction and Survey of the Recent Approaches and Techniques," in 2019 IEEE 9th Symposium on Computer Applications & Industrial Electronics (ISCAIE), 2019, pp. 272–277.
- [18] W. P. Sari and H. Fahmi, "Opinion Mining Analysis on Online Product Reviews Using Naive Bayes and Feature Selection," in 2021 International Conference on Information Management and Technology (ICIMTech), 2021, vol. 1, pp. 256–260.
- [19] G. Badaro et al., "A survey of opinion mining in Arabic: A comprehensive system perspective covering challenges and advances in tools, resources, models, applications, and visualizations," *ACM Trans. Asian Low-Resource Lang. Inf. Process.*, vol. 18, no. 3, pp. 1–52, 2019.
- [20] S. H. Sahir, R. S. Ayu Ramadhana, M. F. Romadhon Marpaung, S. R. Munthe, and R. Watrionthos, "Online learning sentiment analysis during the covid-19 Indonesia pandemic using twitter data," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 1156, no. 1, p. 012011, Jun. 2021.
- [21] Z. Fachrina and D. H. Widyantoro, "Aspect-sentiment classification in opinion mining using the combination of rule-based and machine learning," in 2017 International Conference on Data and Software Engineering (ICoDSE), 2017, pp. 1–6.
- [22] A. Suciati and I. Budi, "Aspect-based Opinion Mining for Code-Mixed Restaurant Reviews in Indonesia," in 2019 International Conference on Asian Language Processing (IALP), 2019, pp. 59–64.
- [23] E. Miranda, M. Aryuni, R. Hariyanto, and E. S. Surya, "Sentiment Analysis using Sentiwordnet and Machine Learning Approach (Indonesia general election opinion from the twitter content)," in 2019 International Conference on Information Management and Technology (ICIMTech), 2019, vol. 1, pp. 62–67.
- [24] H. Wisnu, M. Afif, and Y. Ruldevyani, "Sentiment analysis on customer satisfaction of digital payment in Indonesia: A comparative study using KNN and Naive Bayes," *J. Phys. Conf. Ser.*, vol. 1444, no. 1, p. 012034, Jan. 2020.
- [25] G. A. Buntoro, R. Arifin, G. N. Syaifuddin, A. Selamat, O. Krejcar, and F. Hamido, "The Implementation of the machine learning algorithm for the sentiment analysis of Indonesia's 2019 Presidential election," *IJUM Eng. J.*, vol. 22, no. 1, pp. 78–92, 2021.
- [26] Y. I. Sulistya, "Covid-19 Indonesian Tweet," Kaggle, 2021. [Online]. Available: <https://www.kaggle.com/datasets/yudhaislamisulistya/covid19-tweet-indonesia-positif-dan-negatif/versions/5>. [Accessed: 01-Feb-2022].
- [27] B. Haryanto, Y. Ruldeviyani, F. Rohman, J. D. TN, R. Magdalena, and Y. F. Muhamad, "Facebook analysis of community sentiment on 2019 Indonesian presidential candidates from Facebook opinion data," *Procedia Comput. Sci.*, vol. 161, pp. 715–722, 2019.
- [28] Y. D. Kirana and S. Al Faraby, "Sentiment Analysis of Beauty Product Reviews Using the K-Nearest Neighbor (KNN) and TF-IDF Methods with Chi-Square Feature Selection," *J. Data Sci. Its Appl.*, vol. 4, no. 1, pp. 31–42, 2021.
- [29] P. Desai, J. R. Saini, and P. B. Bafna, "POS-based Classification and Derivation of Kannada Stop-words using English Parallel Corpus," in 2022 3rd International Conference for Emerging Technology (INCET), 2022, pp. 1–5.
- [30] R. Hendrawan and S. Al Faraby, "Multilabel classification of hate speech and abusive words on Indonesian Twitter social media," in 2020 International Conference on Data Science and Its Applications (ICoDSA), 2020, pp. 1–7.
- [31] R. Rosnelly, "The Similarity of Essay Examination Results using Preprocessing Text Mining with Cosine Similarity and Nazief-Adriani Algorithms," *Turkish J. Comput. Math. Educ.*, vol. 12, no. 3, pp. 1415–1422, 2021.
- [32] A. Fauzi, E. B. Setiawan, and Z. K. A. Baizal, "Hoax News Detection on Twitter using Term Frequency Inverse Document Frequency and Support Vector Machine Method," *J. Phys. Conf. Ser.*, vol. 1192, no. 1, p. 012025, Mar. 2019.
- [33] M. Z. Naeem, F. Rustam, A. Mehmood, I. Ashraf, and G. S. Choi, "Classification of movie reviews using term frequency-inverse document frequency and optimized machine learning algorithms," *PeerJ Comput. Sci.*, vol. 8, p. e914, 2022.
- [34] Z. Balani and C. Varol, "Combining Approximate String Matching Algorithms and Term Frequency In The Detection of Plagiarism," *Int. J. Comput. Sci. Secur.*, vol. 15, no. 4, pp. 97–106, 2021.

- [35] G. A. Sandag, "Prediksi Rating Aplikasi App Store Menggunakan Algoritma Random Forest," *CogITo Smart J.*, vol. 6, no. 2, pp. 167–178, 2020.
- [36] P. Chakraborty, F. Nawar, and H. A. Chowdhury, "A Ternary Sentiment Classification of Bangla Text Data using Support Vector Machine and Random Forest Classifier," in *International Conference on Computational Techniques and Applications*, 2022, pp. 69–77.
- [37] H. M. Pandey, P. Tiwari, A. Khamparia, and S. Kumar, "Twitter-based opinion mining for flight service utilizing machine learning," *Inform.*, vol. 43, no. 3, pp. 381–386, 2019.
- [38] W. F. W. Yaacob, S. A. M. Nasir, W. F. W. Yaacob, and N. M. Sobri, "Supervised data mining approach for predicting student performance," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 16, no. 3, p. 1584, Dec. 2019.
- [39] E. Christodoulou, J. Ma, G. S. Collins, E. W. Steyerberg, J. Y. Verbakel, and B. Van Calster, "A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models," *J. Clin. Epidemiol.*, vol. 110, pp. 12–22, 2019.
- [40] K. S. Rawat and I. V. Malhan, "A hybrid classification method based on machine learning classifiers to predict performance in educational data mining," in *Proceedings of 2nd International Conference on Communication, Computing and Networking*, 2019, pp. 677–684.
- [41] S.-S. M. Ajibade, N. B. Ahmad, and S. M. Shamsuddin, "A data mining approach to predict academic performance of students using ensemble techniques," in *International Conference on Intelligent Systems Design and Applications*, 2018, pp. 749–760.
- [42] B. Khotimah, M. Miswanto, and H. Suprajitno, "Optimization of Feature Selection Using Genetic Algorithm in Naïve Bayes Classification for Incomplete Data," *Int. J. Intell. Eng. Syst.*, vol. 13, no. 1, pp. 334–343, Feb. 2020.
- [43] Y. Liu, H. Chen, L. Zhang, X. Wu, and X. Wang, "Energy consumption prediction and diagnosis of public buildings based on support vector machine learning: A case study in China," *J. Clean. Prod.*, vol. 272, p. 122542, 2020.
- [44] A. A. Putra, R. Magdalena, and R. Y. N. Fu'adah, "Klasifikasi Kanker Usus Besar Menggunakan Metode Ekstraksi Ciri Principal Component Analysis Dan Klasifikasi Support Vector Machine," *eProceedings Eng.*, vol. 6, no. 2, 2019.
- [45] P. Birzhandi, K. T. Kim, B. Lee, and H. Y. Youn, "Reduction of training data using parallel hyperplane for support vector machine," *Appl. Artif. Intell.*, vol. 33, no. 6, pp. 497–516, 2019.
- [46] D. Krstinić, M. Braović, L. Šerić, and D. Božić-Štulić, "Multi-label classifier performance evaluation with confusion matrix," *Comput Sci Inf Technol*, vol. 10, pp. 1–14, 2020.

# Classification of Diabetes Types using Machine Learning

Oyeranmi Adigun<sup>1</sup>, Folasade Okikiola<sup>2</sup>  
Department of Computer Science  
Yaba College of Technology  
Lagos, Nigeria

Nureni Yekini<sup>3</sup>  
Department of Computer  
Engineering, Yaba College of  
Technology, Lagos, Nigeria

Ronke Babatunde<sup>4</sup>  
Department of Computer Science  
Kwara State University Malete  
Kwara State, Nigeria

**Abstract**—Machine learning algorithms have aided health workers (including doctors) in the processing, analysis, and diagnosis of medical problems, as well as the detection of disease patterns and other patient data. Diabetes mellitus (DM), commonly referred to as diabetes, is a gathering of a syndrome issue that is portrayed by high glucose levels in the blood over a drawn-out period. It is a long-term illness that is a great threat to humanity and causes death. Most of the existing machine learning algorithms used for the classification and prediction of diabetes suffer from embodying redundant or inessential medical procedures that cause complications and wastage of time and resources. The absence of a correct diagnosis scheme, deficiency of economic means, and a general lack of awareness represent the main reasons for these negative effects. Hence, preventing the sickness altogether through early detection may doubtless cut back a considerable burden on the economy and aid the patient in diabetes management. This study developed diabetes classification using machine learning techniques that will minimize the aforementioned drawbacks in the prediction of diabetes systems. Decision tree classifiers, logistic regression, random forest, and support vector machines are all examples of predictive algorithms that were tested in this paper. 1009 records of data set were obtained from the Diabetes dataset of Abelvikas, Data World. We used a confusion matrix to visualize the performance evaluation of the classifiers. The experimental result shows that the four machine learning algorithms perform well. However, Random Forest outperforms the other three, with a prediction accuracy of 100% and has a better prediction level when compared with others and existing work.

**Keywords**—Machine learning; diabetes mellitus; predictive algorithm; correlation map; confusion matrix

## I. INTRODUCTION

Diabetes is one of the most common and speedily increasing diseases within the world [1] and a serious pathological state in the world. This polygenic disease is a condition in which the body is unable to produce the required amount of internal secretion to keep blood sugar levels in check (National Center for Biotechnology Information, NCBI). In general, a higher risk of diabetes infection is associated with female gender, age over 35, and individuals who are overweight.

The day demands to identify and diagnose this diabetes condition at an early stage cannot be over-emphasized. The diagnosis and analysis of diabetes disease is an important issue

in classification that is required and must be cost-effective, suitable, and valid to be built.

Diabetes mellitus, also known as diabetes, is a metabolic disorder that can result in elevated blood sugar levels (MSD Manual). It is a long-lasting disease that occurs when the pancreas fails to produce enough insulin or when the body fails to properly utilize the insulin that is produced. The insulin in the body system regulates the movement of sugar from the blood into the cells for energy use. Untreated high diabetes blood sugar can cause damage to the critical major organs such as the eyes, and kidneys, heart disease, sudden death which can lead to chronic damage to other organs, etc. [2, 3].

Therefore insulin is a catalyst in the regulation of blood sugar hormones. Hyperglycaemia (high blood sugar) is a common complication of uncontrolled diabetes that resulted in severe damage to nerves and blood vessels of the body's systems [4]. Diabetes is one of the most lethal diseases in the world, but with the introduction of machine learning, there is the potential to find a solution to this pandemic.

The crux of using a machine learning classifier and data mining is to derive knowledge from information stored in the dataset and produce a simple pattern description. A diabetes diagnostic tool using machine learning needs to be developed to predict patients with diabetes to detect the illness early before it is pathetic. Machine-learning algorithms (MLA) identify patterns from statistical quantities of data and feed them into the system to be digitally processed. Much has been achieved in the areas of using machine learning algorithms to solve many challenges in the health sector with the development of technology. Some of these are for the prognosis and/or diagnosis of diabetes for active and accurate decision-making [5]. Therefore, this paper focuses on the application of machine learning techniques to an online dataset to uncover hidden patterns in medical diagnosis and predict diabetes based on the data collected. To ensure that the information obtained from a system built using these techniques is reliable, a Support Vector Machine (SVM) and Random Forest (RF) are proposed for use in the prediction of diabetes in a patient.

It was discovered that there are three major kinds of Diabetes classified into three types: type 1, type 2, and gestational diabetes. Type 1 diabetes is distinguished by a lack of insulin production and necessitates daily insulin administration. Despite the fact that the exact cause of type 1

---

This work was financially supported by Yaba College of Technology in Lagos, Nigeria.



diabetes is unknown, it is unavoidable. The symptoms may appear unexpectedly and are caused by excessive urination (polyuria), fatigue (polydipsia), persistent hunger, loss of weight loss, and vision. Type 2 diabetes (non-insulin-dependent,) may be caused as a result of insufficient insulin in the body and is primarily caused by excess body weight and physical inactivity. The third type is gestational (hyperglycemia), which is defined as having blood glucose levels that are higher than normal but are lower than the conditions of diabetes that occur during pregnancy. This increases the likelihood of complications during pregnancy and childbirth and faces a greater chance of type 2 diabetes in the future too [6].

Patients with diabetes must undergo a series of tests and exams in order to properly diagnose the disease. These tests may include unnecessary or redundant medical procedures that result in complications and a waste of time and resources. Diabetes lowers the standard of living and reduces labor productivity, so the economic cost of the disease far outweighs the direct medical costs within the care sector. The main causes of these negative effects are a lack of a proper diagnosis scheme, a lack of financial resources, and a general lack of awareness. As a result, preventing the illness entirely through early detection will almost certainly reduce the economic burden and aid the patient in diabetes management. The following are the objectives of the study:

- Develop the Diabetes prediction system using a decision tree classifier, logistic regression, random forest, and support vector machine.
- Evaluate and compare the developed system and the performance of each algorithm in the ensemble of algorithms based on sensitivity, specificity, and accuracy.

The study is organized into five sections. Section I introduces the study by discussing the keywords briefly as well as the study's objectives Section II explains various related works in the field of diabetes type prediction Section III describes the study's methodology in detail. Section IV discusses the results of the algorithms. Section V concludes the study with recommendations for additional research.

## II. RELATED WORKS

Several researchers have made contributions to fields where diabetes was predicted. Diabetes has a significant economic impact on society, and it is the most expensive chronic disease. The author [7] addressed the fact that majority of diabetes patients are asymptomatic, which leads to delayed standard clinical laboratory examinations that create large health datasets over a lifetime. They looked at machine learning algorithms to help with diabetes screening via routine laboratory tests, using data from 62,496 patients' lab tests. The following classifications were used; artificial neural networks, Bayes naïve, K-nearest neighbor, random forest, regression models, and support vector machines. In detecting diabetes, the artificial neural network model outperformed the others. Based on clinical data processing, computer processing has been used to identify diseases [8]. Knowledge extraction from data to aid

decision-making by experts is a movement in the next generation of intelligent health systems [9].

The author [10] sought to develop effective models for predicting early gestational diabetes mellitus (GDM). The seven variables and 73 variables datasets were used to create models that predicted early GDM in different situations. In early pregnancy, ML models predicted GDM with high accuracy and were developed and tested in the Chinese population. The study [11] also employed ML and classification algorithms, with Logistic Regression providing the highest accuracy of 96 percent. Also, [12] carried out the use of random forest, KNN, Nave Bayes (NB), and J48 to develop diabetes analysis and prediction. The researchers used two datasets: PIDD (Pima Indian Diabetes Dataset) and 130 US hospital diabetes data sets. The developed system achieved 93.62 percent accuracy in the case of PIDD and 88.56 percent accuracy for a large dataset of 130-US hospitals. For large dataset analysis, the NB and J48 prediction algorithms were found to be superior. The author [13] reviewed diabetes types and treatments, as well as some emerging issues that may arise, and listed physical activities that will lead to healthy lifestyles.

Furthermore, [14] presented diabetes prediction based on big data from healthcare communities using various machine learning algorithms. Using SVM for classification and K-means for clustering, the developed system used an effective strategy for detecting diabetes disease earlier. The study [15] implemented a decision tree algorithm to predict diabetes. The experiments were carried out on the Pima Indians diabetes database, and the results achieved an accuracy of 87 percent. However, low sample sizes result in poor accuracy. The system that was developed can be used to predict or diagnose other diseases in the same family. In a similar vein, [16] used the most recent records of 13,309 Canadian patients aged 18 to 90 years, as well as their laboratory data. They developed predictive models using Logistic Regression and Gradient Boosting Machine (GBM) techniques and compared them to others such as Decision Tree and Random Forest. The GBM and LR models outperform the other two models. In this experiment, [17] proposed two machine learning classification algorithms, Fine Decision Tree and Support Vector Machine, which are used to detect diabetes at an early stage. When compared to the Fine Decision Tree algorithm, the SVM classification algorithm achieved a high percentage of accuracy.

The author [18] applied random forest, decision tree, and neural network in their study to predict diabetes mellitus with an accuracy of about 81 percent. The Pima Indians diabetes dataset from the UCI machine learning repository was used. The study [19] proposed using classification algorithms to predict diabetes. On a number of criteria, three machine learning classification algorithms were researched and assessed. According to the experimental findings, the Naive Bayes classification algorithm has an accuracy rate of 76.30 percent.

In [20] the author proposed the use of the Pima Indians diabetes dataset, using Decision Tree, K-Nearest Neighbors, Support Vector Machine, and Random Forest to predict diabetes at various stages and compare the performance of

different classification techniques. While [21] presented a Unified Framework for Diabetes Prediction Based on Machine Learning. Six machine learning classifications for predicting diabetes and various evaluation criteria were used to investigate the performance of these classification techniques. The analysis results show that Naïve Bayes achieved the highest performance than the other classifiers, obtaining the F1 measure of 0.74. According to [22] in the prediction of diabetes using the classification algorithms. Naive Bayes, Multilayer Perceptron, and IBK algorithms were used. The Naive Bayes algorithm shows 100% accuracy compared to IBK 88% and Multilayer perceptron 88%.

Research work made by [23] developed a machine learning-based framework for detecting type 2 diabetes in electronic health records. The system created a semi-automated framework based on machine learning. A data-informed framework for identifying subjects with and without T2DM from EHR was proposed using machine learning and feature engineering. The author in [24] created an Ontology-based Diabetes Management system, a computer-based system that assists physicians in correctly diagnosing diabetes mellitus disease in patients. They used the Bayesian Optimization technique to boost prediction accuracy. Similarly, [25] developed a medical expert system for diabetes diagnosis, a diabetes ontology with 9 sub-classes, and a web-based application with web service architecture. With test data from 65 patients, an overall consistency rate of 90.7 percent was achieved. The author [26] demonstrated diabetes detection at an early stage using a computational intelligence fuzzy hierarchical model capable of performing early detection and identifying someone's susceptibility to DM. The model's accuracy is 87.46 percent. A number of techniques have been proposed over the years for the prediction of diabetes types. The comparison of diabetes techniques in Table I shows their performance and limitation. Four different classifiers will be used, and because Random Forest excels at working with non-linear data, the prediction will be more accurate and stable, with improved performance.

TABLE I. COMPARISON ANALYSIS OF EXISTING TECHNIQUES

S/N	Author(s)	Strategy	Performance %	Limitations
1	Aishwarya & Vaidehi (2019)	LR, RRF, RF.	96.	more time spent on their synthesis.
2	Minyechil et al (2019)	Random Forest, KNN, Naïve Bayes, and J48	93.62	time-consuming processes.
3	Quan Zou et al (2018)	D, RF, NN	80.8	could not predict the type of diabetes.
4	Zheng et al (2017)	KNN, Naïve Bayes, DT, RF, SVM, & LR	95	The model distinguishes patients with and without type 2 Diabetes Mellitus
5	Aiswarya et al (2015)[27]	DT & Naïve Bayes	J48 76.9, NB 79.5	not precise and a general conclusion for diabetes

### III. METHODOLOGY

The importance of early diagnosis of diabetes mellitus to the life expectancy of the patient suffering from it cannot be over-emphasized. Early diagnosis will mean that, based on certain biological features found in the medical history of the patient, there is a predictive test. This section focuses on how predictive analysis of machine learning is used to predict the diabetes status of a patient accurately. Therefore, to develop and implement a diabetes recommendation prediction system, the proposed model employs machine learning techniques.

#### A. Predictive Analysis

This section illustrates the analysis of the proposed system and how the system that was designed works and is a feasible alternative to the existing one. The data used in this paper was collected from the dataset of Abelvikas, Data world. The data collected were subjected to different types of pre-processing, as will be addressed in subsequent sections to improve the system's performance. The proposed model implements the classification model with the highest accuracy level. These algorithms include Logistic Regression, Decision Tree, Random Forest, and Support Vector Machine Classifiers. Fig. 1 shows the block diagram for the proposed model.

1) *Data acquisition*: This research was carried out using the dataset of Abelvikas, Data world. The dataset has multi-class problems of diabetes which separate it into individuals who have tested either negatively or positively (type1, type2, and normal) to diabetes. The dataset consists of 1009 total instances with eight attributes to provide adequate data from training after pre-processing requiring the removal of certain entries. Entries included Age (years), BS Fast (mmol/L), BS pp (mmol/L), Plasma R mmol/L, Plasma F (mmol/L), and HbA1c (mmol/L). The data collected from the Abelvikas, Data World Database was shown in the Table II.

2) *The pre-processing stage*: This handles inconsistencies in data to improve accuracy and precise outcomes. This dataset has missing values for a few selected attributes like Glucose level, Blood Sugar, and HBA1C because these attributes cannot have values of zero. The dataset is then scaled to normalize all values. Correlation is an amount of context between characteristics. It is a real number value that denotes the degree of significance between 0 and 1 and a negative value indicates an inverse relationship, while a direct relationship is indicated by a positive value. Fig. 2 shows the correlation map of the proposed model.

3) *Training & classification*: ML algorithms require training data to achieve the objective. This training dataset will be analyzed by the algorithm, which will then classify the inputs and outputs before analyzing it again. A sufficiently trained algorithm will effectively memorize all of the inputs and outputs in a training dataset. The prediction model consists of the best machine learning model after implementing different models, and the best was taken and deployed for application. The output of each model is taken to the next stage for testing. In training the classification algorithms and constructing the model, the steps taken were to

import the modules and dataset as a data frame and get insights from the dataset. From the entire data set, a feature set containing the first seven attributes is extracted, and the output set is extracted, which is the product of the prediction and the whole set is split into a 7:3 ratio train set and test set.

4) *Testing*: After the model was built, testing data validate to make accurate predictions. This is to confirm that the ML algorithms were trained effectively to evaluate the prediction models created.

5) *Evaluation*: Assessing the performance of the model using different metrics is integral to this research work. Based on the result from the test stage, the model was evaluated based on classification accuracy and specificity. A classification metric was employed to evaluate the developed model. There are four types of outcomes that could occur when performing classification predictions.

a) True positives happen when you predict that an observation belongs to a certain class and it turns out to be correct.

b) True negatives occur when you predict that an observation will not belong to a class and it actually does not belong to that class.

c) False positives happen when you assume an observation belongs to a class when it doesn't.

d) False negatives occur when you incorrectly predict that observation does not belong to a class when it does.

The results are frequently plotted on a confusion matrix. After making predictions based on the test data and then classifying each prediction as one of the four possible outcomes described above, the matrix was generated.

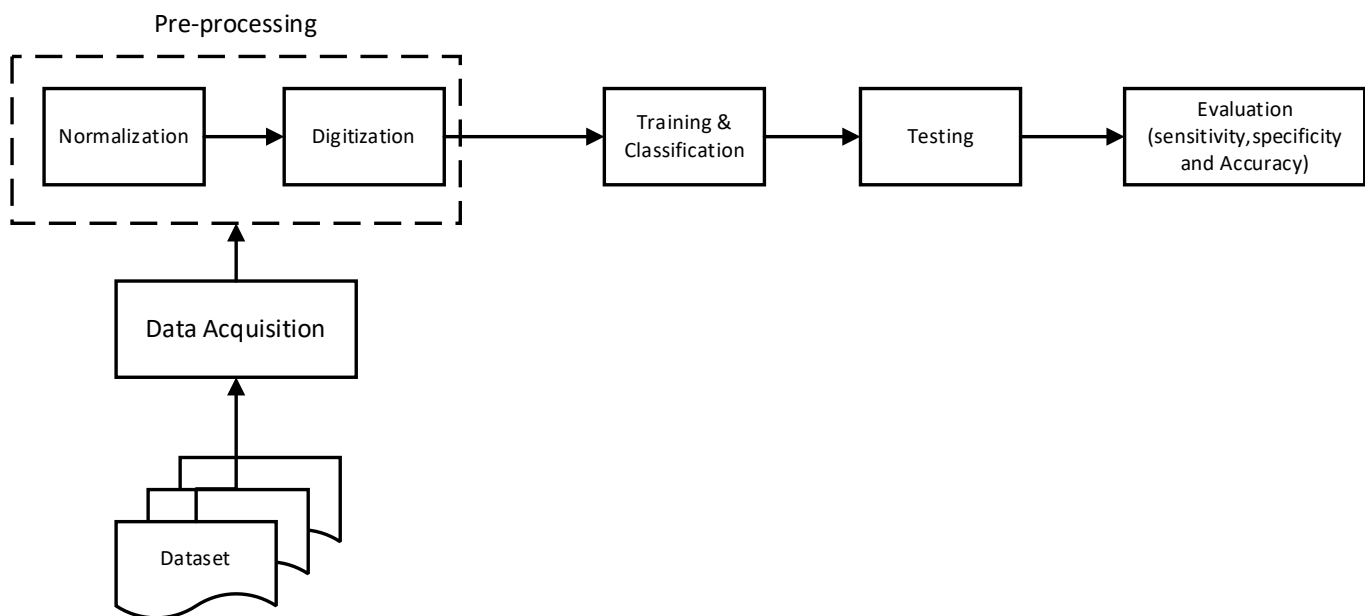


Fig. 1. Proposed Model for the Research.

TABLE II. DATABASE FILE REPRESENTATION

S/N	Field	Type	Range
1.	Age	Integer	21 – 81
2.	Blood Sugar in fasting	Real	0 – 54
3.	Blood Sugar after a meal	Real	4.2 - 8.1
4.	Plasma Glucose in fasting	Real	3.9 - 9.1
5.	Plasma Glucose	Real	7.9 - 13.1
6.	Glycated hemoglobin (HBA1C)	Real	28 – 69
7.	Type	String	0 – 255
8.	Class	Boolean	1 - Diabetic, 0 - Non-Diabetic

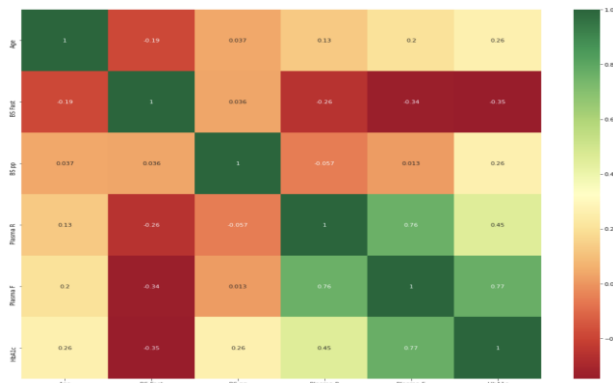


Fig. 2. Correlation Map.

### B. Modelling Methods

In training the classification algorithms and constructing the model, the following steps were taken (see Fig. 3).

Step One: Import the modules and dataset as a data frame

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import warnings
warnings.filterwarnings('ignore')
warnings.filterwarnings('ignore', category=DeprecationWarning)
df = pd.read_csv("../content/Diabetestype.csv")
```

Fig. 3. Code Snippet to Import Dataset.

Step Two: Get insights from data

The Insight derived from the dataset is shown in Table III

TABLE III. INSIGHTS GOTTEN FROM DATASET

	Age	BS Fast	BS pp	Plasma R	Plasma F	HbA1c	Type	Class
0	50	6.8	8.8	11.2	7.2	62	T1	1
1.	31	5.2	6.8	10.9	4.2	33	N	0
2.	32	6.8	8.8	11.2	7.2	62	T1	1
3.	21	5.7	5.8	10.7	4.8	49	N	0
4.	33	6.8	8.8	11.2	7.2	62	T1	1

Key: T1=Type1, N=Normal

Step Three: Specify features and test sets

A training set containing the first seven attributes of the data set is extracted and the test set which is the eighth attribute is also extracted, which is the product of the prediction and the entire dataset is split into a 7:3 ratio train set and test set.

Step Four: Train prediction model

The prediction model consists of the best machine learning model after implementing different models, and the best was taken and deployed for application. The output of each model is taken to the next stage for testing.

Step Five: Test model

The test set is used to assess the prediction models that have been created. This step is carried out four times to ascertain consistency.

Step Six: Evaluate

From the result of the test stage, the model is evaluated based on classification accuracy and specificity. The Table IV shows the accuracy of the four models based on these parameters.

1) *Prediction methods for diabetes*: The following machine learning strategies are used for comparative analysis of the diabetes predictive model. Classifiers include logistic regression, decision trees, random forests, and support vector machines.

a) *Logistic Regression (LR)*: It is another supervised learning classification algorithm that models the relationship between a categorical response variable and its covariates. It computes probabilities using a logistic function, which is the accumulative logistic distribution, to assess the association between a categorical dependent variable and more than one independent variable. It is another probabilistic-based statistical model used in machine learning to solve classification problems. The logistic regression model uses the sigmoid function to predict the probability of outcomes of positive and negative class and can be derived from a sigmoid function obtained below,

$$P = \frac{1}{1+e^{-a-bx}} \quad (1)$$

where P = probability, a and b = parameter of Model.

b) *Decision Tree Algorithm*: A DT is one of the supervised machine learning algorithms that employ the classification regression trees algorithm, which can handle both classification and regression. It aids decision-making by generating a decision-tree-like model in which data is continuously split according to a specific parameter. There are two types of units in the tree: decision nodes and leaves. The data is split at the decision nodes, and the final decisions or outcomes are at the leaves. To solve classification and regression problems, the algorithm generates decision trees from training data. The classification error rate is defined as the proportion of the training set that does not belong to the most common class:

$$\text{Entropy (S)} = \sum_{i=1}^n -P_i \text{Log}(P_i) \quad (2)$$

where  $P_i$  is the percentage of the training set from the  $i$ th class in the region.

c) *Random Forest (RF)*: It is one of the machine learning prediction algorithms. It lends itself better to the ensemble approach. It is capable of handling large datasets with ease. Random Forest is an ensemble classifier made up of many decision trees, with the ensemble implying that it employs multiple machine-learning algorithms to achieve predictive performance. It outperforms others in terms of diabetes mellitus prediction.

The following is the algorithm:

- 1 Create an N-Tree bootstrap sample using the input data.
- 2 Grow an unpruned regression for each bootstrap sample by splitting the node from all predictor nodes. Predictors select the best split from the input variables.
- 3 Predict new data by aggregating N-Tree predictions.

The random forest formulas are given below using Gini Index formulae for classification.

$$\text{Gini Index} = 1 - \sum_{i=1}^n (P_i)^2 \quad (3)$$

It measures total variance across  $i$ th classes that true positives and true negatives. It should be noted that the Gini index is a measure of node purity that has a small value if all of the  $P_i$  are close to zero or one.

#### Support Vector Machine

A Support Vector Machine (SVM) is a type of supervised classification algorithm that has been widely and successfully applied to text classification tasks. This helps with regression and classification tasks and can work with multiple variables. This algorithm effectively performs nonlinear classification and also maps the inputs into a high-dimensional feature space that is used for classification, detection, and regression.

Step 1: Identify the appropriate hyperplane.

Step 2: Following the first step, the second step is to maximize the distances between neighboring data points.

Step 3: Insert a feature  $z = x^2 + y^2$ . It implies that SVM can solve such a problem.

Step 4:-Use an SVM classifier to classify the binary class.

SVM formulae are derived from the equation of hyperplane function to obtain the below,

$$W^* = \arg_w \text{Max} \frac{1}{\|W\|_2} [\text{Min } Y_n | W^T (\phi(x) + b)] \quad (4)$$

Where  $\arg_w \text{Max}$  is an acronym for arguments of the maxima, which are simply the locations of a dynamic array domains where a function's particular value is maximized. The inner phrase  $[\text{Min } Y_n | W^T (\phi(x)+b)]$  essentially indicates the shortest distance between two points and the closest point to the decision boundary.

#### IV. RESULTS AND DISCUSSION

This section aims to get acquainted with results obtained after performing various activities on the dataset obtained from the dataset of Abelvikas, Data World. Fig. 4 shows the registration page for the patient.

##### A. Results

The implementation tools used in this research are Python programming language, google collaboratory, and libraries containing algorithms used for artificial intelligence development, and Anaconda houses a large amount of these libraries. Fig. 5 depicts the recent patient, daily added patient and diabetes rate charts. Fig. 6 shows the disease diagnosis and report.

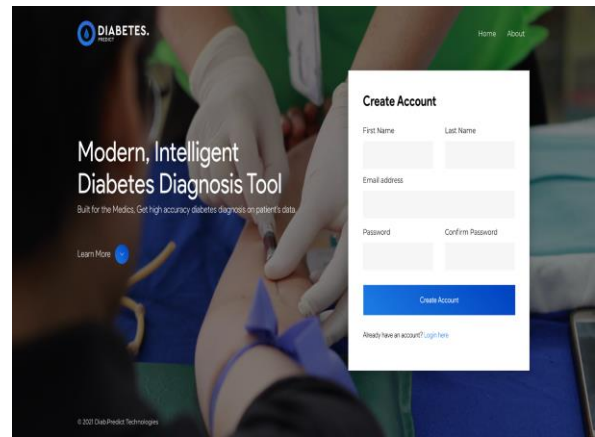


Fig. 4. Landing Page and Registration Page.

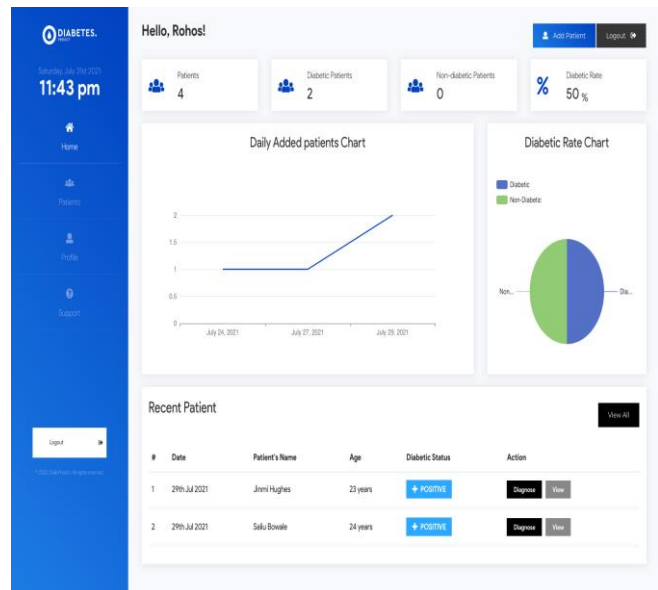


Fig. 5. Dashboard Page.

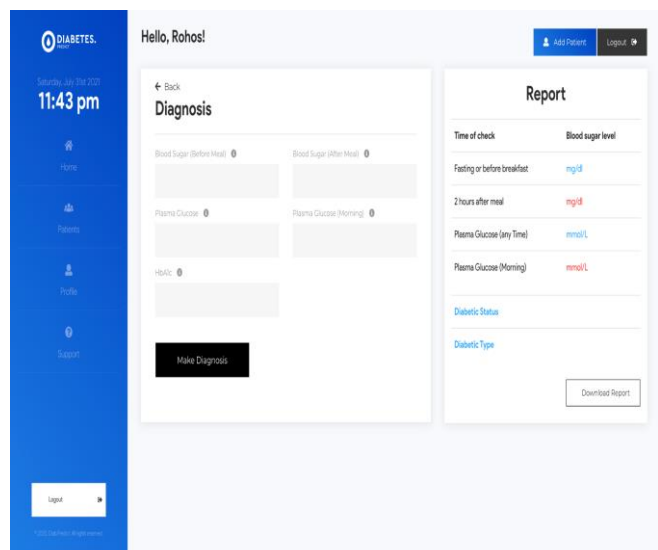


Fig. 6. Diagnose Patient Page.

1) *Performance metrics*: The classifiers we used were then applied to the dataset individually and ran five iterations to ensure that the results obtained from the average of each implementation of a particular algorithm are accurate. Also, these tests were done on randomly selected samples of the dataset to avoid the problem of overfitting. Various parameters were used to evaluate the system, but for this research, three performance indexes were used: Sensitivity (SE), Specificity (SP), and accuracy, as shown in equations (5)–(7). True positives (TP) and true negatives (TN), as well as the false positives (FP) and false negatives (FN).

$$(SE) = \frac{\text{no\_of\_predicted\_true\_positive}}{\text{true\_positive} + \text{false\_negative}} \quad (5)$$

$$(SP) = \frac{\text{no\_of\_predicted\_true\_negative}}{\text{true\_negative} + \text{false\_positive}} \quad (6)$$

$$\text{Accuracy} = \frac{\text{number\_of\_correct\_predictions\_}}{\text{total\_number\_of\_prediction}} \quad (7)$$

The prepared model was integrated into a Python Web Framework, and Flask Framework and hosted on a server for testing. To test the solution, random records from the dataset were used, and an average of the following was calculated for each algorithm. From the above list, it is shown amongst our Ensemble of algorithms why the Random forest algorithm was chosen as the eventual algorithm used for the implementation of this work, as it has the highest average accuracy among the four algorithms.

2) *Confusion matrix evaluation*: A confusion matrix is also referred to as a contingency table or error matrix, used to visualize the performance of a classifier, it's a good way of evaluating a good effective classification model. This means that the high performance of any classification model can be visualized in its confusion matrix having a strong main diagonal shown in Fig. 7.

3) *Implementation of confusion matrix*: The confusion matrix was implemented for each algorithm in the ensemble of algorithms leading to the results are shown in Table IV, while Table V shows the confusion matrix for Classification Models using Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF), and Decision Tree (DT).

The above shows the result of the confusion matrix for classification algorithms with 70% training data and 30% testing data of 1009 records. The result yielded the Table V below.

$$y = \frac{1}{N} \sum_{i=1}^5 Xi \quad (8)$$

where  $y$  is the mean,  $Xi$  is the result of the confusion matrix and the  $i$ -th attribute value of the no of iterations.

$$y = \frac{1}{N} \sum_{j=1}^3 yi \quad (9)$$

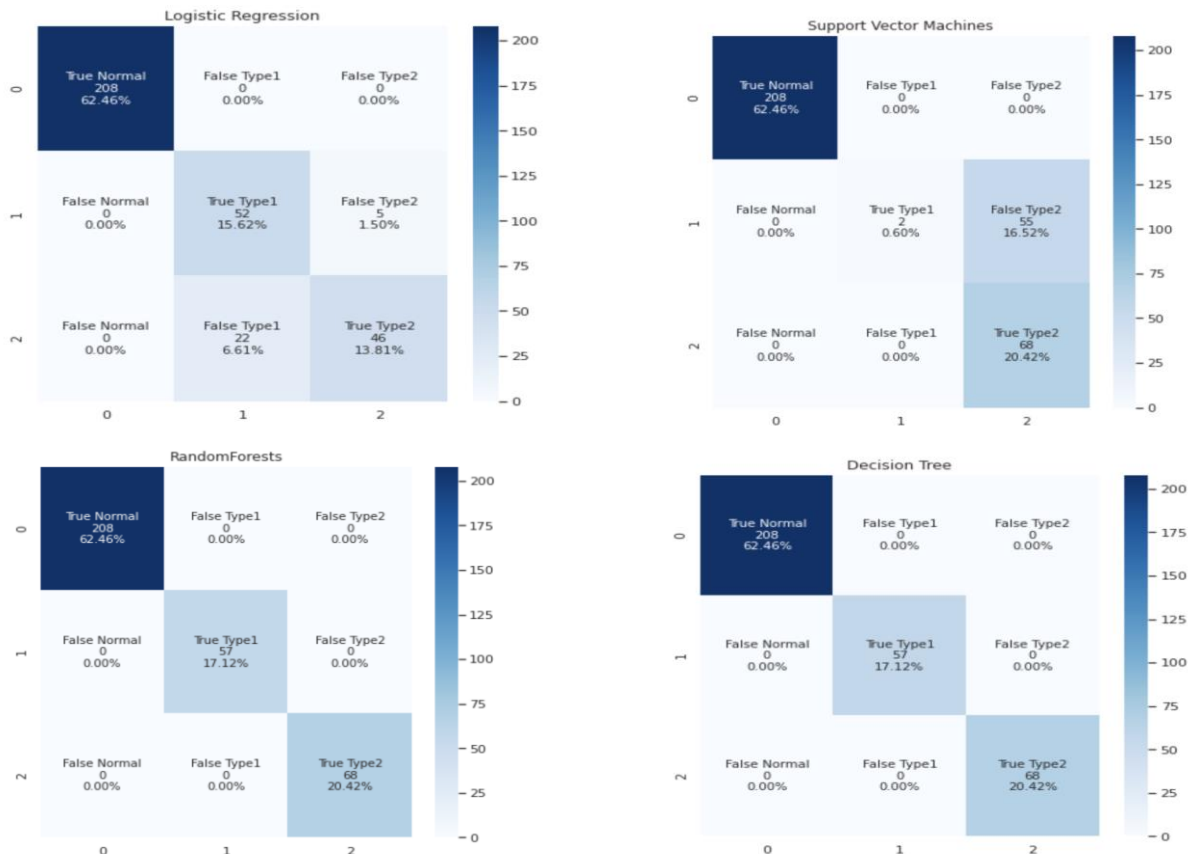


Fig. 7. Confusion Matrix for Classification Models using (i) LR (ii) SVM (iii) RF (iv) DT.

TABLE IV. PERFORMANCE OF THE STUDIES CLASSIFICATION MODEL USING NORMAL, TYPE 1 AND TYPE 2 DIABETES

Algorithm	Class	1 <sup>st</sup> Iteration	2 <sup>nd</sup> Iteration	3 <sup>rd</sup> Iteration	4 <sup>th</sup> Iteration	5 <sup>th</sup> Iteration	Mean Accuracy
LR	Normal	0.990099	0.955446	0.955446	1.000000	0.940594	0.997030
	Type 1	0.940594	1.000000	0.940594	0.940594	0.995050	0.955426
	Type 2	0.960396	0.955446	1.000000	0.980100	0.980100	0.954436
RF	Normal	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000
	Type 1	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000
	Type 2	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000
DT	Normal	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000
	Type 1	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000
	Type 2	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000
SVM	Normal	0.995050	0.831683	0.826733	1.000000	0.856436	0.998020
	Type 1	0.856436	1.000000	0.851485	0.851485	0.995050	0.842417
	Type 2	0.831683	0.826733	1.000000	0.840796	0.840796	0.840436

Key: LR=Logistic Regression, RF= Random Forest, DT=Decision Tree, SVM = Support Vector Machine  
after the confusion matrix are similar to the classification model results. Examining the confusion matrix revealed the same similarity. Table VII shows the mean average score of the algorithms.

where  $y$  is the mean average,  $y_i$  is the mean of the result of the confusion matrix, and the  $i$ -th attribute value of the number of classes.

TABLE V. CONFUSION MATRIX DATA FOR CLASSIFICATION MODELS USING (I) LR (II) SVM (III) RF (IV) DT (%)

Classifier	Normal	Type1	Type2
LR	62.46	22.23	15.31
SVM	62.46	0.60	36.94
RF	62.46	17.12	20.42
DT	62.46	17.12	20.42

In Fig. 8 the use of an ensemble of algorithms aids data mining in determining the most effective algorithm that can be used to generate an effective model. The accuracy report obtained from multiple tests shows that the random forest and decision tree algorithms on our dataset proved to be better prediction algorithms than the other algorithms. The results were compared to the results of works of literature. The Table VIII demonstrated that the developed system's accuracy was higher than [28] accuracy of 91.32 percent because RF excels at working with non-linear data, constructing multiple decision trees, and merging them to produce a more accurate and stable prediction with improved performance.

### B. Discussion

Diabetes has recently become one of the leading causes of death in humans. Diabetes is becoming more common every year for a variety of reasons, including poor eating habits, and the prevalence of unhealthy foods. Diabetes detection early on can help with clinical management decision-making. We have employed numerous measures of evaluation throughout this research to determine and quantify the performance of each algorithm in our ensemble of algorithms, which comprises the Logistic Regression algorithm, Decision Tree, Random Forest, and Support Vector Machine Classifier algorithms, all these algorithms were tested on the diabetes dataset of Abelvikas in five iterations, and the result of the test gave a model that we eventually used for the implementation. However, with all these algorithms it was important to realize which was the most effective of all them, and this was achieved by getting an accurate reading of each and, including algorithm over five iterations. An average of the accuracy reading from each algorithm was used as a measure to determine the eventual algorithm that was used to form our model (Fig. 9), which turned out to be the Random forest and Decision tree Algorithms. From Table VI, the outcomes of the average of the accuracy tests on each algorithm are displayed, this also includes the specificity accuracy and sensitivity accuracy as well as the classification accuracy. When we compare the values in Tables V and VI, we see that the classification results

TABLE VI. COMPARISON OF RESULTS OF DIFFERENT CLASSIFIERS

Metrics Average	Logistic Regression	Decision tree Classifier	Random Forest	Support Vector Machine
Accuracy (%)	92	100	100	83

TABLE VII. THE MEAN AVERAGE SCORE OF THE ALGORITHMS

Metrics Average	Logistic Regression	Decision tree Classifier	Random Forest	Support Vector Machine
Sensitivity	0.921705	1.0	1.0	0.692391
Specificity	0.980548	1.0	1.0	0.933251
Accuracy	0.968964	1.0	1.0	0.893624



V. CONCLUSION

This study compares and evaluates the performance of four machine learning algorithms in the classification of diabetes. The Abelvikas datasets from the Data World repository are used to train and test the system. For diabetes classification, a host of machine learning models have been applied with 1009 instances and eight critical variable features were extracted and identified: age, blood sugar in fasting, blood sugar after a meal, plasma glucose in fasting, plasma glucose, glycated hemoglobin, type, and class. The results of the analysis revealed that the Random forest and Decision tree models were the most accurate in predicting diabetes. The system developed ensures a stable prediction. As a result, the models can be more effectively applied to other diseases. A combination of algorithms, rather than just the most performant algorithm in the ensemble, may be more beneficial in the future.

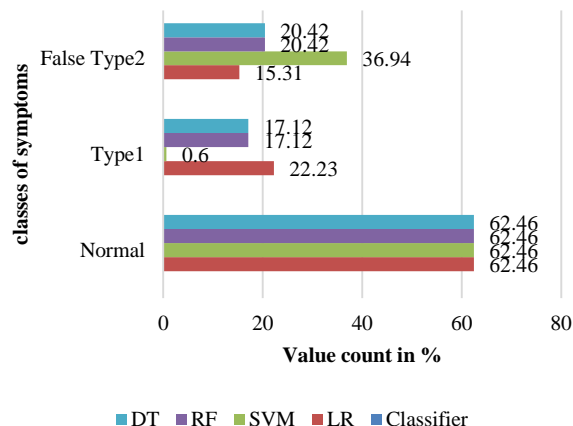


Fig. 8. Confusion Matrix Data for Classification Models.

TABLE VIII. RESULTS COMPARISON TABLE

Author	Model / Method	Dataset Used	% Accuracy
1. Deepti & Dillip. 2018	Naive Bayes & SVM	PIMA Indian Diabetes dataset	76.3%
2. Radha, et al. (2014)	C4.5	A hospital repository	86%
3. Song et al.. (2017)	ANN	Small undefined number of data	74.8%
4. Rashid, & Abdullah, 2016	Decision Tree	A hospital repository	75.5%
5. Afrand, (2012)	Combination of Classifier algorithms	A hospital repository	91.3%
6. Adidela (2012)	Fuzzy ID3 and Estimation maximization algorithm	A private hospital Repository	91.3%
7. Developed System	LR RF DT SVM	Abelvikas, Data world	92% 100% 100% 83%

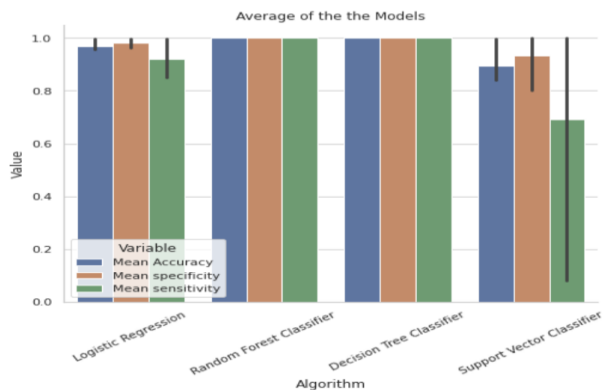


Fig. 9. Average Accuracy of the Models.

REFERENCES

- [1] W.H.O. "About diabetes". World Health Organization, 2014
- [2] A. Krasteva,., V. Panov., A. Krasteva., A. Kisselova, and Z. Krastev, "Oral cavity and systemic diseases-Diabetes Mellitus." Biotechnol. Biotechnol. Equip. 25, 2183-2186. Doi: 10.5504/BBEQ.2011.
- [3] Mahmud, S M Hasan, Hossin , Md Altab, Md. Razu Ahmed, Sheak Rashed Haider Noori, Md Nazirul Islam Sarkarm. "Association for Computing Machinery. ACM ISBN 978-1-4503-6582-6/18/08 DOI: https://doi.org/10.1145/3297730.3297737n, 2018.
- [4] Angela Betsaida B Laguipo. "COVID-19 could trigger diabetes in Healthy people". News Medical Life Science, 2020.
- [5] Lee, Yong-ho, Ban,g Heejung and Kim Dae Jung, "How to establish Clinical Prediction Model", Journal List Endocrinol Metab, 2016.
- [6] P. Samant., and R. Agarwal, "Machine learning techniques for medical diagnosis of diabetes using iris images. Computer Methods and Programs in Biomedicine". 157, 121–128. DOI: https://doi.org/10.1016/J.CMPB.2018.
- [7] Glauco Cardozo , Guilherme Brasil Pintarelli , Guilherme Rettore Andreis ,Annelise Correa Wengerkievicz Lopes, and Jefferson Luiz Brum Marques, "Use of Machine Learning and Routine Laboratory Tests for Diabetes Mellitus Screening. Hindawi BioMed Research International Volume, pp1-14, 2022.
- [8] M. E. Hossain., A. Khan, M. A. Moni, and S. Uddin, "Use of electronic health data for disease prediction: a comprehensive literature review," Transactions On Computational Biology And Bioinformatics, vol. 18, no. 2, pp. 745–758. 2021.
- [9] De Silva K., N. Mathews, H. Teede et al. "Clinical notes as prognostic markers of mortality associated with diabetes mellitus following critical care: a retrospective cohort analysis using machine learning and unstructured big data," Computers in Biology and Medicine, vol. 132, article 104305. 2021.
- [10] Wu et al. (2021) Early Prediction of Gestational Diabetes Mellitus in the Chinese Population via Advanced Machine Learning, The Journal of Clinical Endocrinology & Metabolism, Vol. 106, No. 3, e1191–e1205 doi:10.1210/clinem/dgaa899 Clinical Research Article.
- [11] Aishwarya, Mujumdar, V. Vaidehi "Diabetes prediction using machine learning algorithms International Conference on Recent Trends in Advanced Computing", ICRTAC, 2019.
- [12] Minyechil, Alehegn, Rahul, J. & Dr. Preeti, M. "Diabetes Analysis And Prediction Using Random Forest, KNN, Naive Bayes, And J48: An Ensemble Approach".International Journal of Pure and Applied Mathematics, Volume 118 No. 9,871-878m, 2019.
- [13] Nail, Rachel, and Suzane Falck, "An overview of diabetes types and treatments".https://www.medicalnewstoday.com/articles/323627, 2020.
- [14] Vijayakumar, Kavin Prasad Arjunan, Manivel Sivasakthi, Karthikeyan Lakshmanan "Diabetes Prediction By Machine Learning Over Big Data From Healthcare Communities", International Research Journal of Engineering And Technology(Irjet)E-Issn: 2395-0056volume: 06 Issue: 04| Apr2019.

- [15] Thomas, Jesia , Anumol Joseph, Irene Johnson, Jeena Thomas, "Machine Learning Approach For Diabetes Prediction" International Journal of Information Systems and Computer Sciences: Volume 8, No.2, 2019.
- [16] Hang Lai, Huaxiong Huang, Karim Keshavjee, Aziz Guergachi, and Xin Gao "Predictive models for diabetes mellitus using machine learning techniques" Lai et al. BMC Endocrine Disorders <https://doi.org/10.1186/s12902-019-0436-6>, October 2019.
- [17] H. R. Divakar, D Ramesh, B R Prakash "An Ontology-Driven System to Predict Diabetes with Machine Learning Techniques," International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, pp 4005-4011, Vol-9 Issue-2, 2019.
- [18] Quan Zou, Kaiyang Qu, Yamei Luo, Dehui Yin, Ying Ju, Hua Tang. "Predicting Diabetes Mellitus with Machine Learning techniques". <https://dx.doi.org/10.3389%2Ffgene.2018>.
- [19] Deepti S. & Dilip S. S. "Prediction of Diabetes using Classification Algorithms". *International Conference on Computational Intelligence and Data Science*, pp 1578-1585 (ICCIDIS 2018),
- [20] Farooqui, Ritika, and Tyagi, "Prediction Model for Diabetes Mellitus Using Machine Learning Techniques. International Journal of Computer Sciences and Engineering Open Access Research Paper Volume-6, Issue-3 E-ISSN: 2347-2693, 2018.
- [21] Mahmud, S M Hasan, Hossin , Md Altab, Md. Razu Ahmed, Sheak Rashed Haider Noori, Md Nazirul Islam Sarkarm. "Association for Computing Machinery. ACM ISBN 978-1-4503-6582-6/18/08 DOI: <https://doi.org/10.1145/3297730.3297737n>, 2018.
- [22] K. Nandhini.M, "Prediction of diabetes using Classification algorithms," International Journal of Science, Engineering and Management, vol. 2, no. 12, pp. 287-291, 2017.
- [23] Zheng, Tao, Wei Xie , Liling Xu , Xiaoying He, Ya Zhang, Mingrong You, Gong Yang , You Chen "A machine learning-based framework to identify type 2 diabetes through electronic health ", 2017.
- [24] F. M. Okikiola, O.S. Adewale, A.M. Mustapha, A.M. Ikotun, O.L. Lawal "A framework for Ontology-based diabetes diagnosing system using bayesian optimization technique. <https://doi.org/10.51406/jnset.v17i1.1906>. 2018.
- [25] Sakorn Mekruksavanich."Medical expert system based ontology for diabetes disease diagnosis"IEEE International Conference on Software Engineering and Service Science (ICSESS) pp 383-389, 2016.
- [26] Rian Budi, Lukmanto E.Irwansyah The Early Detection of Diabetes Mellitus (DM) Using Fuzzy Hierarchical Model," *Procedia Computer Science* Volume 59, 2015, Pages 312-319.
- [27] I. Aiswarya., S. Jeyalatha., S. Ronak.," Diagnosis of Diabetes using classification mining techniques". *International Journal of Data Mining & Knowledge Management Process (IJDKP)*, Vol. 5, No. 1, pp. 1-14, 2015.
- [28] D. R. Adidela, "Application of fuzzy ID3 to predict diabetes." *Int J Advanced Computer Math Sci* 3.4.541-5., 2012.

# Criteria and Guideline for Dyslexic Intervention Games

Noraziah ChePa<sup>1</sup>, Nur Azzah Abu Bakar<sup>2</sup>, Laura Lim Sie-Yi<sup>3</sup>

Human-Centred Computing, School of Computing  
Universiti Utara Malaysia, Sintok, Kedah  
Malaysia

**Abstract**—The utilization of game-based interventions is growing as a result of technological advancements, and it has shown to be effective in the treatment of dyslexia and other medical conditions. Games are typically viewed as activities having the essential components of challenge, incentive, and reward. Games were originally created for pleasure, and they can make dyslexic teaching and learning more enjoyable and exciting. Although there are numerous applications available for treating dyslexic children, the inclusion of games and their standards in those applications has not yet been established. Therefore, there is a need for a standard design guideline to be formulated in establishing a guideline for designing and developing games specifically for dyslexic children. This article proposes a design guideline for dyslexic intervention games. Two methods have been employed which are interviews and systematic literature reviews (SLR) to discover the characteristics of dyslexic games. The first set of the criteria was developed through interviews with the stakeholders who are directly associated with dyslexic children. Scopus, the ACM digital library, EBSCO-host, Wiley, and Web of Science (WOS) are the five primary databases used in SLR. 50 articles out of the 551 that were early screened from the five primary databases are qualified to be studied based on the criteria. Only 23 publications could be selected for the study after further screening, which led to the creation of a second set of criteria. These two sets of criteria are thoroughly analyzed, combined, and formulated as a guideline which comprises of four main categories; device and platform, interface, game features, and gameplay. The guideline consists of guidance to be used for designing and developing Dyslexic therapy games with the purpose of assisting Dyslexic children to read. The guideline is believed to be beneficial to many parties especially the educational game developers, therapists, and educationist who are dealing with intervention for Dyslexic children. This study is aligned and significant to Sustainable Development Goals (SDG) three and four, Good Health and Well-being and Quality Education respectively.

**Keywords**—*Dyslexic therapy games; game-based intervention; specific learning disorder; guideline for dyslexia games; dyslexia intervention*

## I. INTRODUCTION

Dyslexia refers to specific learning disorder (SLD) that involves difficulty in reading due to problems identifying speech sounds and learning how they relate to letters. It is listed as a mental disorder in the International Classification of Mental Disorders and the Diagnostic and Statistical Manual of Mental Disorders [1]. It resulted from an unexpected phonological deficit [2]. In other words, children with dyslexia

have low ability in decoding and spelling. Often a dyslexic child will have trouble connecting the sound made by a specific letter or deciphering the sounds of all the letters together that form a word. Dyslexia have increasingly been found to be the most common learning disability accounting up to 80% of the learning-disabled population in general [3].

As of 2020, it is reported that between 5 to 20% of the world population struggled to read due to dyslexia [4]. In 2017, the Dyslexia Association of Malaysia reported 10% of the school age children in Malaysia were affected by the disorder. The percentage shows an increase from 2014 in which 53,685 students with learning disabilities have been involved in formal education and from that total, 0.03% or 1,681 students have been involved in the dyslexia classroom programme [5]. Earlier study conducted by Socio-economic and Environmental Research Institute of Penang has identified 9.4% of children in Grade One elementary schools in Penang as having learning difficulties, and 92.3% of these children were found to have severe reading disabilities [6]. It is one dyslexic case in every 20 students, compared to one down-syndrome case in every 600 people or one spastic case in every 700 people [7].

Despite the emphasis on literacy difficulties, dyslexia would appear to include a wide range of symptoms including poor short-term memory, dyscalculia, visual impairment, speech disorders, and poor motor control as well as emotional difficulties such as poor self-esteem, clinical depression, chronic anxiety and conduct disorders. The deficits in these keystone academic skills lead to poor academic achievement and they tended to lag far behind in age and intellectual ability from their peers. The impact can change at different stages in a person's life. It can seriously affect a person's self-esteem [8]. Dyslexia students sometimes feel dumb, frustrated, lonely, humiliated, and academically less competent than they really are. They may get very frustrated and are at risk of developing mental health problems such as anxiety and depression. Despite all these, dyslexia are not related to intelligence or lack of desire to learn.

Although dyslexia are typically thought of as learning difficulty with educational consequences, there is increasing evidence that dyslexia are also associated with health difficulties. Auto immune disorders, allergies, autism and schizophrenia are common amongst families where there are learning disabilities [9]. Furthermore, [10] observed that dyslexic children with the most severe symptoms of fatty acid deficiency (rough skin, dry skin and hair) have the most severe reading, spelling and short-term memory difficulties.

There is lack of consensus on how dyslexia should be diagnosed or treated. One systematic review [11] revealed that traditional special educational methods have limited impact on dyslexic children. Improvements through intensive reading interventions yield small to moderate effects overtime, and there appears to be a subset of 25% of problem readers who do not respond to special education. In the absence of satisfactory remediation through traditional special educational methods, there have been several alternative treatments offered for dyslexia. These include biofeedback, hypnotherapy, music therapy, visual occlusion therapy, the neural organization chiropractic technique, primary reflex therapy and Dyslexia Dyspraxia Attention Treatment (DDAT) exercises.

Past studies on dyslexia[12]–[15] utilized games, either for identification or intervention purposes, as discussed further in Chapter 2. Games add more fun and excitement to teaching and learning. Quite a few games have been developed as an alternative treatment for dyslexia. These include the board games (e.g. Zingo Sight Words, Scrabble Junior, Brainbox ABC, Monopoly Junior and Alphabet Lotto) as well as the online or digital games that run either on IOS or Android or both platforms (e.g. Draw Something, Hanging with Friends, Anagram Scramble, ABCya, Chicktionary, Boogle Bash, Knoword and Word Whomp). Different strategies are used in these games such as draw out a given word, spell a complex word, create ambiguous word to puzzle others, make out words from a given set of letters, find word while beating the time allotted, and complete words by conjecture based on the word's definition and first letter.

Designing games for the dyslexic needs careful consideration on their special needs in order to maximize their learning experience and overcome their difficulties. Several criteria for designing and developing an effective game for dyslexia intervention have been discussed in the literature, however, they are yet to be formalized into a standard guideline. To date, very limited number of standard guidelines exists and none of these focus specifically on games [16]–[18].

The absence of systematic guideline in designing and developing therapy games for dyslexic children is key issue to be solved in this study. To date, very few systematic guidelines have been developed. The [17] guidelines focus on early detection of dyslexia and is not meant for therapy. Whereas the [18] guidelines focus on learning reading but their emphasis is on user interaction with the application, i.e., how to design interfaces that are affective for the dyslexics. The guideline excludes important game elements such as goals, rewards, challenge and feedback. The author [16] developed a guideline for dyslexic games; however, the coverage is limited to user interface aspects of the games. Many other criteria discussed in the literature are not yet formalized into a guideline and thus, the process of designing and developing therapy games for dyslexia is time consuming as the developers need to gather and analyse the criteria from various sources.

Considering the addressed issues, this article aims to identify the criteria and formalize the guideline to be used in the design of dyslexic games. This will serve as a reference that would be beneficial for the developers of games or tools for dyslexic children.

## II. DYSLEXIA AND INTERVENTION

The current state of dyslexia research can be characterized by the distinction of scientists in groups of protagonists of a visual versus a phonological/auditory deficit on the one hand and in groups of protagonists of a low, basic level versus a higher-level deficit on the other hand. A lot of contradictory results and theories posed the question about specificity and homogeneity of different deficits in dyslexic individuals. The model of [19] provides an integration of perceptive and cognitive deficits based on a common temporal processing deficit, which can be analyzed on a low, basic level and/or on a higher complex level of performance. Over the last 35 years, there has been a great deal of research focused on finding the most effective methods for treating dyslexia. This body of knowledge is complex, in part because although all individuals with dyslexia have a similar problem namely, difficulty in reading, they have heterogeneous characteristics, and depending on the child's developmental level, the demands of reading and the required skills are quite different [20].

The paper [21] suggests that difficulties in literacy acquisition for dyslexics are due to lack of phonological awareness, problems to recognize words and understand spelling rules, visual errors in spelling, letter and word confusion with similar-sounding words and omissions of words, parts of words and individual letters and sounds. In other words, their literacy skills are at word-level reading and spelling [22]. The dyslexics have difficulties in identifying phonemes and the exchanging of letters occurs very often during the spelling process; they also often mixed-up the letters of 'b-d', 'u-n', 'm-w', 'g-q', 'p-q', and 'b-p' [7], [23]. Evidence of their great difficulties in writing, poor skill of spelling, oral and written vocabulary and also weak in arranging the content of the compositions is also found in [24]. Besides, previous studies also found that children with dyslexia are significantly slower at naming colours, digits and letters, thus suggesting that children with dyslexia have persistent, and unexpectedly severe problems in naming speed for any stimuli [25].

The study [7] summarizes the difficulties in spelling, reading and writing faced by most dyslexic children as in Table I.

TABLE I. PROBLEMS FACED BY DYSLEXIC CHILDREN

Problem	Description
Spell	Confusion in identifying letters such as: <ul style="list-style-type: none"><li>• m – w; y – g – j; u – n; m – n; c – e; p – q; h – n; b – d</li></ul> Confusion in the letter sound such as: <ul style="list-style-type: none"><li>• t – h; f – v; s – h; r – l</li></ul>
Read	Reversal in the word such as: <ul style="list-style-type: none"><li>• Batu – tuba</li><li>• Gula – lagu</li></ul> Reversal in the sentence such as: <ul style="list-style-type: none"><li>• Pada masa yang sama – dapa masa yang masa</li></ul> Confusion between Malay and English word such as: <ul style="list-style-type: none"><li>• Jam – jem; cat – cat</li></ul>
Write	Difficulty holding a pencil; cannot write according to the line provided; tends to write words fads.

At a basic level of spelling, learning to represent sounds with letters requires a two-way mapping between phonology and written symbols, and it is here that difficulties will first be encountered by children with phonological deficits [26]. They need to acquire knowledge of the relationship between sounds and letters which requires them to be familiar with phonological representations and the correspondence between phoneme and grapheme. In the next step, they need to segment the target word into its salient sounds and then represent these sequentially with symbols [26]. What makes this exacerbated is the fact that spelling, unlike reading, is difficult to use context. In her study, [27] found a significant difference in the nature of the spelling errors in dyslexic children compared with a control group. They made 'phonetically unacceptable' errors that may not be recognized as the word because of a lack of phonetic similarity. This implies that the dyslexic children may have not developed phonological representation but use letter naming strategies to spell phonologically regular words.

Regarding reading, [28] suggests that for sight word reading to develop, learners must acquire and apply knowledge of the alphabetic system. According to [29], lexical processing, or the ability to recognize words quickly and accurately, is a symbol of skilled reading. In the context of Malay language, a study by [30] reveals both syllable awareness and phoneme blending are significant predictors of word recognition; when the readers have inefficient syllable segmentation, oversimplification of syllables, insufficient grapheme-phoneme knowledge and inefficient phonemic code assembly they will make errors in reading.

#### A. Dyslexia Interventions

Various intervention methods or treatments have been used to manage the literacy and cognitive abilities for children with dyslexia [31]. Most research used experimental designs [32]–[34]. There were a few studies that applied multimedia training in their intervention program. The researches [35] and [33] utilised computer-assisted training while [36] used video games in their remedial intervention. The study [37] adopted Magnocellular deficit theory in their treatment plan. A few studies employed multisensory approach [33], [38]–[42]. This approach uses graphics and strong colors to make associations between shape, letters, words and numbers that relate to the same topic, and involves techniques for linking eyes, ears, voice, and hand movements to symbolic learning. The approach taken is to try and engage as many sensory receptors in the learning process as possible, since it is argued that on many occasions, children with learning difficulties appear to have stronger sensory receptors over their non-dyslexic peers.

The majority of the studies targeted language or literacy components such as writing skills, reading skills, word and alphabet mastery as the outcome of the study [32], [33], [39], [40]. Many also carried out intervention or training based on specific impaired cognitive function such as visual-motor intervention and working memory training [35], [37], [38], [43], [44]. In other words, they have chosen a specific difficulty to be treated in their intervention.

To date, there are limited modules or intervention programmes being conducted in Malaysia for children with dyslexia. In fact, there is currently no standardized module for

dyslexia class in Malaysian public school [31]. Traditional methods in teaching these children to overcome their difficulties in the classrooms were found not to be encouraging and were not successful in overcoming their difficulties in reading [40]. Thus, traditional methods have a limited impact on dyslexic children [11], [45].

#### B. Digital Game-Based Interventions

Recent decades have witnessed the increasing use of digital interventions with game-like components. Initially developed for entertainment, a "game" is generally considered to be an activity with the key features of challenge, motivation, and reward. Digital game-based interventions have been found to have practical effects in addressing the main barriers of access and engagement in the healthcare domain, particularly for the young [46]. Evidence from neuropsychological research further suggests the therapeutic value of digital game-based interventions in depression therapy. Positive game-playing experiences are claimed to have triggered the release of hormones such as endorphins and striatal dopamine that are responsible for feelings of pleasure and well-being [47]. Interventions using games have also been used in therapy or rehabilitation sessions of various other illnesses such as brain injury [48], cerebral palsy [49] and upper limb injury [50].

Past studies utilized games for diagnosis or early detection of dyslexia [17], [51]–[53] as well as for intervention [12]–[14], [54]–[58]. Other studies [56], [59] use gamification approach by utilizing some game elements (such as scores or rewards) in non-game context or the so-called application which ranges from desktop, mobile and web-based applications.

As evidenced in [55] and [60], games for dyslexia often implement a collection of activities that are connected by a common visual theme and can be played independently of each other. On the other hand, a smaller number of games are built around a rich narrative in which the student progresses through a story by solving language-related activities. For example, iLearnRW [13] draws together these two alternatives and provides a common interface and metaphor for the visualization of learning progress. The game is built around characters rather than a fixed story; although there is no guarantee for infinite re-playability, this has improved player's engagement. Each character in the game represents a group of language difficulties that the student will practice every time he/she initiates an activity related to them. Following an initial interaction with a new character, the student earns that character as a friend, who is displayed on a social network-like interface. The design of iLearnRW also considered the need for personalized intervention as some students learn faster than the others or prefer playing sessions at different length.

#### C. Existing Guideline

The study [17] derived a set of guidelines to design an optimal tablet game for 5-year-old children. The guidelines, called DYSL-X, focus on early detection of dyslexia and was used to develop Diesel-X, a game about a robot dog character named Diesel, which must fight against a gang of criminal cats. Diesel-X contains three games, which require players to know letters. However, there is no further information about the validation DYSL-X. The author [18] proposed an affective

interaction design (IxD) model to facilitate reading for the dyslexics. Their study emphasized on the need to use proper interface when designing an application for the dyslexics, by taking into account the affective attributes of the dyslexic children. In [16] proposed a guideline for dyslexic game design which focused on user interface aspects, i.e., usage of font, colour, navigation, consistency, interaction and game type. Recent studies [16], [52], [53], [61]–[64] revealed multiple criteria which are further analysed and formalized into the guideline developed throughout this study. The detailed criteria from individual studies are presented in Section IV.

### III. METHODOLOGY

Due to its suitability and compatibility with the scope of study, a Design Science Research methodology [65] is adapted in designing the research framework. Research methodology is divided into four main phases; awareness of problems, identification of criteria, formulation of game guideline, and evaluation as illustrated in Fig. 1.



Fig. 1. Research Methodology.

#### A. Awareness of Problems

The first stage of conducting this study is about identifying issues and problems that lead to the formulation of solutions. Issues and problems are identified through content study and personal communications involving stakeholders in treating Dyslexic children. The absence of systematic guideline in designing and developing therapy games for Dyslexic children is key issue to be solved in this study.

#### B. Identification of Criteria

To identify the game criteria, combination of systematic interview and Systematic Literature Review (SLR) have been employed. Interview has been conducted involving eleven respondents among teachers and students at Dyslexia Incubator, School of Computing, UUM. Respondents involved are among the content experts, five teachers, and potential users who are the children diagnosed with Dyslexia.

SLR is conducted by adopting PRISMA approach in conducting SLR [66]. This study has focused on five main databases which are Scopus, ACM digital library, EBSCOhost, Wiley, and Web of Science (WOS). Four main processes that have been carried out in the searching process are identification, screening, eligibility, and data extraction and analysis as shown in Fig. 2.

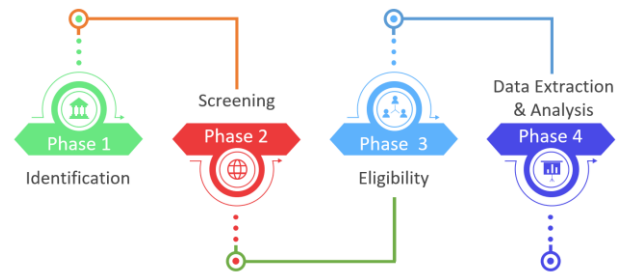


Fig. 2. Four Main Stages of Reviewing Literatures.

1) *Identification*: The first phase is about determining the keywords to be used for searching. In this context, keywords related to reading disabilities and the affected group, which are the children are basically relied on. The use of game intervention in treating the affected age group are also used in the searching. All keywords that have been used specifically for the database involved are listed in Table II.

TABLE II. KEYWORDS AND SEARCHING INFORMATION STRATEGY

Databases	Keywords used
Scopus	TITLE-ABS-KEY ((dyslexia OR ("reading disabilities" AND (child* OR kid))) AND (game OR "game intervention"))
ACM	[[Publication Title: dyslexia] OR [[Publication Title: "reading disabilities"] AND [[Publication Title: child*] OR [Publication Title: kid]]] AND [[Publication Title: game] OR [Publication Title: "game intervention"]] OR [[Abstract: dyslexia] OR [[Abstract: "reading disabilities"] AND [[Abstract: child*] OR [Abstract: kid]]] AND [[Abstract: game] OR [Abstract: "game intervention"]] OR [[Keywords: dyslexia] OR [[Keywords: "reading disabilities"] AND [[Keywords: child*] OR [Keywords: kid]]] AND [[Keywords: game] OR [Keywords: "game intervention"]]
EBSCOhost	TI ((dyslexia OR ("reading disabilities" AND (child* OR kid))) AND (game OR "game intervention")) OR AB ((dyslexia OR ("reading disabilities" AND (child* OR kid))) AND (game OR "game intervention"))
Wiley	"(dyslexia OR ("reading disabilities" AND (child* or kid))) AND (game OR "game intervention")" in Title, Abstract and keyword
WOS	TS=((dyslexia OR ("reading disabilities" AND (child* OR kid))) AND (game OR "game intervention"))

2) *Screening*: For screening of the relevant articles, several conditions for inclusion and exclusion have been defined. Type of literature, language, and subject area are among the criteria that have been included. Criteria and its eligibility terms are defined in Table III.



TABLE III. THE INCLUSION AND EXCLUSION CRITERIA

Criterion	Eligibility	Exclusion
Literature type	Journal (research articles)	Journals (systematic review), book series, book, chapter in book, conference proceeding
Language	English	Non-English
Subject Area / Categories	Computer Science	Other than Computer Science

The searching also focused on literatures with empirical data such as journal articles, research articles, and review articles. To prevent difficulties of translation, only English articles are included. For relevance, articles with related to the focus are selected which are related to the use of intervention games that involve children with reading difficulties. For this phase, 10 articles have been removed.

3) *Eligibility*: The main focus of this phase is to identify the eligible articles to be included in the study based on the criteria explained earlier. To achieve this, the identified articles are reviewed and thoroughly analyzed. Focus is given on the targeted objectives. To identify the details, abstracts will be reviewed before the articles will be analyzed thoroughly. Fig. 3 illustrates the processes involved by adapting PRISMA method.

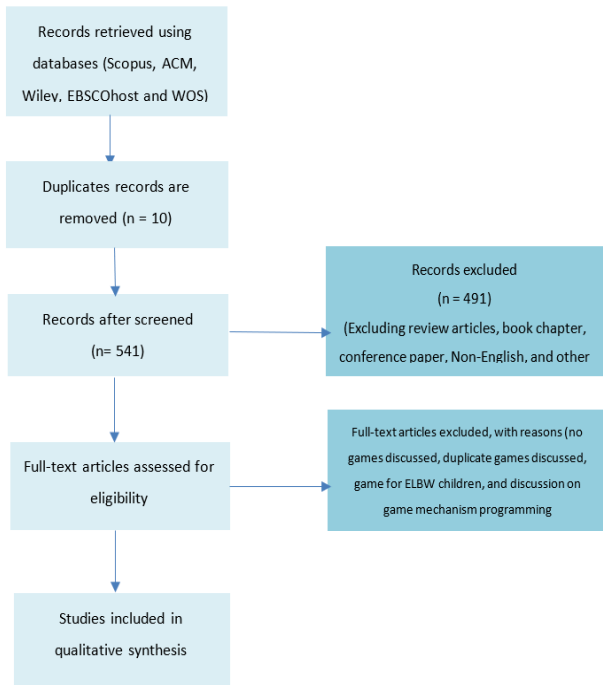


Fig. 3. Adaption of PRISMA Approach in Selecting the Articles.

From the process, there are 551 articles have been early screened from five main databases to be further reviewed. From this total article, there are 10 duplicate articles and have been removed for the next process. 541 articles were further screened based on the relevancy to the criteria. However, only

50 articles are eligible to be analyzed. After further screening, there are only 23 articles which are significant to be included in the study.

4) *Data extraction and analysis*: This phase is focusing on extracting and analyzing key details from the 23 articles that have been chosen. Various types of data are extracted from these articles. They are proponents of the criteria, year of study, game title, and the criteria of the game. The list of the extracted criteria is covered in findings section of Section IV. Deliverables of this phase are two sets of the criteria for Dyslexic game.

C. Formulation of Game Guideline

List of criteria gathered from SLR and interview are thoroughly analyzed and compared in formulating the guideline. Set one of the criteria is extracted from SLR, while set two of the criteria is gathered from interview. Processes involved in formulating the guideline are shown in Fig. 4.

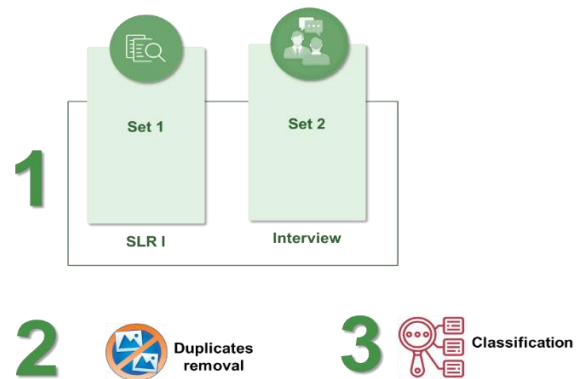


Fig. 4. Processes Involved in Formulating the Guideline.

Second process involved the removal of duplicates of the criteria by analyzing similarities exist in both sets. Removal also involved for cases when some criteria are using different words but are referring to the same thing. The last step is classification, where all criteria are mapped into four main categories; device and platform, interface, game feature, and gameplay. The outcome of these processes is covered in Section IV.

D. Evaluation

To ensure that the proposed guideline is correctly formulated and meets its specification, both verification and validation are conducted in evaluating the proposed guideline. For verification, expert review has been conducted involving five experts from different areas; game development experts, educationist, and counsellors. For validating the guideline, a prototype of Dyslexic game, namely *DysRedia* is developed. The outcome of the evaluation and its detail discussions are covered in Section IV.

IV. THE PROPOSED GUIDELINE

The proposed guideline is formulated based on the two sets of criteria gathered and extracted from systematic interview and the existing studies through systematic literature review as



explained in Section III. The first set of criteria are extracted from 23 related articles from SLR. The extracted criteria are sorted based on its publication years. The listed criteria are carefully analyzed by focusing on similarity of the criteria proposed by its proponents. Duplicates are removed. The criteria are then classified into four main categories: device and platform, interface, features, and gameplay. There are five, eleven, thirteen, and sixteen criteria classified into four categories respectively.

Second set of the criteria are gathered from interview session involving teachers and students. Table IV depicts 30 criteria gathered from interview that have been categorized into four main categories.

TABLE IV. CLASSIFIED CRITERIA GATHERED FROM INTERVIEW

Category	Criteria	
<b>Device &amp; platform</b>	1. Mobile app 2. Tablet	3. Touch-based
<b>Interface</b>	1. Simple interface 2. Font & background: F8 (white font, red background)	3. Font type: comic 4. Font size: 16 5. Small caps
<b>Features</b>	1. Audio 2. Video 3. Still picture 4. Animation	5. Background music 6. Letter with phonic 7. Letter arrangement: keyboard design 8. Attractive images
<b>Gameplay</b>	1. Exercises 2. Different difficulty levels 3. Competition 4. Hints 5. Help 6. Tutorial 7. Rewards, more rewards at higher levels	8. Scoreboard 9. Replay 10. Auditory feedback 11. Different categories 12. Real images 13. More exercise in the same difficulty level 14. Levels arranged alphabetically

There are three, five, eight, and fourteen criteria have been classified into four categories respectively. Fig. 5 shows number of criteria and its category that have been successfully classified.



Fig. 5. Number of Criteria Acquired from Two Method and Classified into Four Categories.

These two sets of criteria are then combined based on its category. For Device and Platform, five criteria from SLR and three criteria from interview are combined producing only five criteria, as three of them are redundant and have been removed. The final five criteria are shown in Fig. 6.

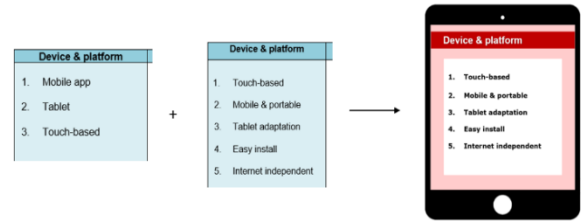


Fig. 6. Criteria for Device and Platform Category.

For Game Features category, eleven and five criteria identified from SLR and interview respectively have been combined and produced a guideline of 13 criteria by removing three redundant criteria. The final thirteen criteria are shown in Fig. 7.

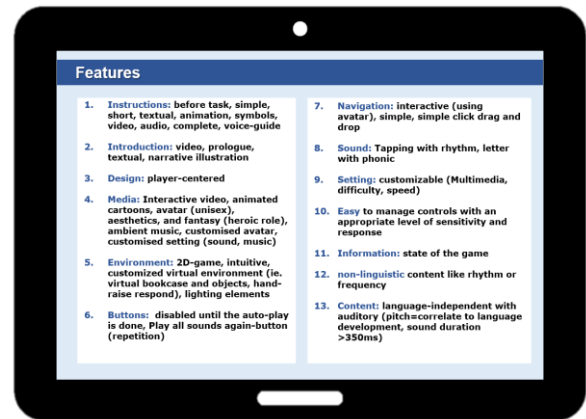


Fig. 7. Criteria for Game Features Category.

For Interface category, combination of 13 and eight criteria identified from SLR and interview respectively, have produced a final set of eleven criteria. Ten redundant criteria have been removed. Terms referring to the same criteria such as ‘small caps’ suggested by interview and ‘lower case’ suggested by SLR are combined. We decided to use ‘lower case’ in the criteria as suggested by most of the proponents. A final version of eleven criteria for Interface category is shown in Table V.

For Gameplay category, sixteen and fourteen criteria identified from SLR and interview are combined and produced a final set of 18 criteria. Twelve redundant criteria have been removed. Criteria using different terms but are referring to the same criteria are also removed. A proposed guideline for Dyslexic intervention games comprising of all final combined criteria are illustrated in Fig. 8.

TABLE V. CRITERIA FOR INTERFACE CATEGORY

Interface	
1. Simple interface	2. Font size: 16, 18, 18-26, minimum font size of 14 points
3. Font colour: grey scale in font (10%), text in black using a mono-spaced, dark colour on light background (Suggest cream color background, white font, red background)	4. Font type: lower case, Arial typeface, typeface Courier, font style Verdana, OpenDyslexic font, Helvetica, Comic san)
5. Background: grey scale in the bg (90%) crème/black color pairs, text in black on creme background, plain background, brilliant/bright colors	6. Layout: Fixed , Unobstructed views, playful, line spacing (1.4), paragraph spacing (2), column width (77 character/line), column width not wider than 60 characters per line, consistent, character spacing (+7%), Child friendly color (unique to induce positive emotion) and shape (round)
7. Figures: simple geometric	8. Icon: 3D
9. Graphic/visual: simple, consistent, appealing, easy to interpret, large and touchable, attractive, search-like, non-related linguistic, letter arrangement: keyboard design, real images	10. Character: cute, children friendly
11. Fantasy-themed setting	

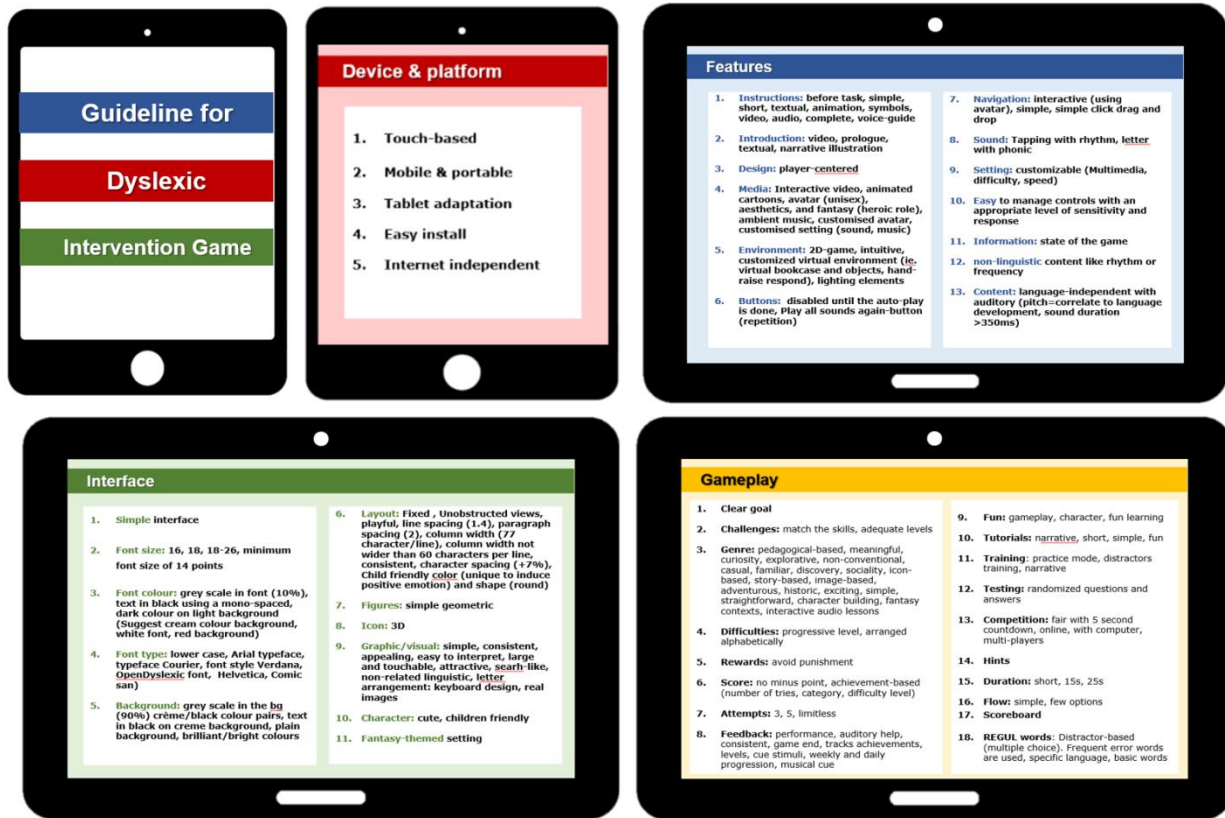


Fig. 8. The Proposed Guideline.

1) *Evaluation of the proposed guideline:* The proposed guideline has been evaluated through two methods; expert review and prototyping. Verification is conducted before validation. Five experts were involved in verifying the proposed models in ensuring that the proposed guideline confirms its specification. Table VI listed five experts involved in reviewing the proposed guideline.

TABLE VI. EXPERTS BACKGROUND

Expert	Field of expertise	Qualification & designation	Years of experience
Exp1	Special education	Dyslexia teacher	2 years
Exp2	Counselling	Trainee	4 years
Exp3	Computer system and network	PhD, Lecturer	More than 20 years
Exp4	Interaction design	PhD, Lecturer	20 years
Exp5	Counselling	Trainee	4 years

There are eight components that have been used in reviewing the proposed guidelines; clarity, visibility, comprehensive, evolutionary, flexibility, accuracy, understandability, and effectiveness. Clarity is meant to evaluate whether the guideline is clearly presented. There are four constructs used to measure the clarity of the guidelines; (C1)-the whole design guideline for Dyslexic game is clearly presented, (C2)-the categories in the design guideline for Dyslexic game are defined clearly, (C3)-the elements in the design guideline for Dyslexic game are defined clearly, and (C4)-all relations between the categories and elements are clearly presented. For measuring visibility, three constructs have been used; (V1)-the design guideline for Dyslexic game is visible to be followed, (V2)-the guides involved can be followed easily, and (V3)-the design guideline for Dyslexic game can be a guide by developers to solve related tasks of design and development. Fig. 9 shows the mean score for each construct for clarity, visibility, comprehensive, and evolutionary.

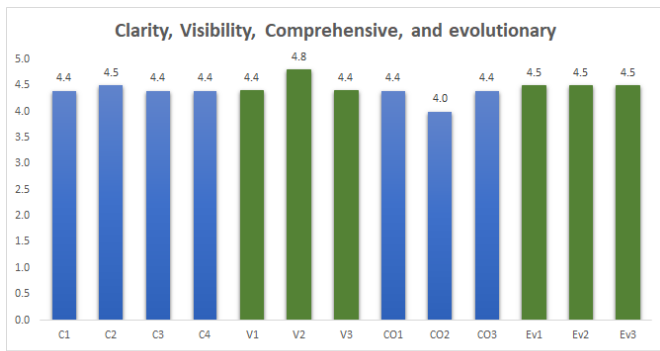


Fig. 9. Expert's Review on Clarity, Visibility, Comprehensive and Evolutionary.

To measure the completeness of the guideline, three constructs have been used; (CO1)-the whole design guideline for Dyslexic game is defined completely, (CO2)-the design guideline for Dyslexic game covers all related elements, and (CO3)-the relations between the categories and elements are sufficient. While for measuring the ability of the guideline to evolve, three constructs are used; (Ev1)-the design guideline for Dyslexic game is dynamic, (Ev2)-the design guideline for Dyslexic game allows additional factors in the future, and (Ev3)-the design guideline for Dyslexic game provides opportunity for improvements.

Mean score for all constructs in measuring clarity, visibility, comprehensive, and evolutionary are considered high (more than 2.5). It can be concluded that all experts agreed that the proposed guideline is clear, visible, complete, and able to evolve. Another four aspects used in measuring the guidelines are flexibility, accuracy, understandability, and effectiveness. There are three, four, two, and three constructs used for each aspect respectively. Constructs used to measure flexibility are; (F1)-the design guideline for Dyslexic game is flexible to be edited, (F2)-the design guideline for Dyslexic game is adaptive to changes, and (F3)-the design guideline for Dyslexic game is generalizable enough to be applied for other related tasks.

While for measuring how accurate the guideline is, two constructs are used; (A1)-the design guideline for Dyslexic game is presented correctly and (A2)-all categories and elements factors are labelled correctly. Four constructs are used to measure understandability; (U1)-the whole design guideline for Dyslexic game is easy to understand, (U2)-The label of each category is understandable, (U3)-the label of each factor is understandable, and (U4)-adhering to the design guideline for Dyslexic game is easy. The last aspect is the efficiency of the proposed guideline. There are three constructs used; (Ef1)-the design guideline for Dyslexic game can guide in the development of engaging Dyslexic games, (Ef2)-adhering to the design guideline for Dyslexic game will improve the engagement of Dyslexic games, and (Ef3)-adhering to the design guideline for Dyslexic game will improve usability in Dyslexic games. Fig. 10 shows the mean score for each construct for flexibility, accuracy, understandability, and effectiveness.

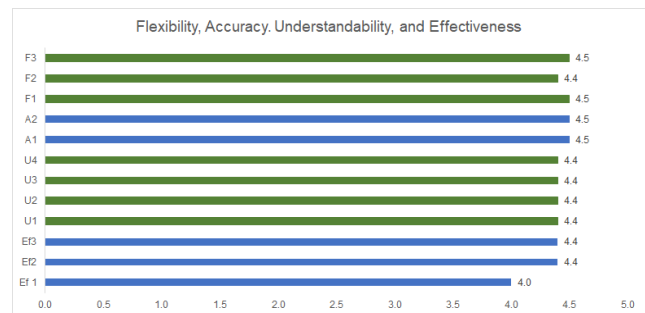


Fig. 10. Expert's Review on Flexibility, Accuracy, Understandability and Effectiveness.

Mean score for all constructs in measuring flexibility, accuracy, understandability, and effectiveness are considered high (more than 4.0). It can be concluded that all experts agreed that the proposed guideline is accurate, flexible, can be easily understood, and effective. Experts have also given their overall review of the guideline. They are particular on visual aspects of the guideline especially on interface elements as suggested by Expert 1, Expert 4, and Expert 2.

- E3 : Customizable font size
- E1 : Font style: = use the written letter 'a' not 'a'
- E2, E5 : Font style = choices between capital letter and small case

Font size is important element involving Dyslexic children. Three experts suggested font size to be customizable in the game. Expert 1 is consistent with what is proposed in the guideline on the font type, as San Serif font should be used. Expert 2 and 5 are also suggested that players should be given options to choose between capital letters or small letters to be used in the game. Other elements also attracted Expert's attention, for example background music (pleasant to children's ear) and hints (disclosing one letter). They prefer options to be given in turning on background music. While for hints, they suggested that hints only to be given when they win the game. Too many hints will discourage learning among players as suggested by Expert 1.



- E1 : Background music= on/off button
- E1 : Gain hint through winning, they will keep use hint if too many hint
- E4 : The guideline should be printed in bigger size.
- E1 : Application Icon = attractive

Experts are concerned about the look of the proposed guideline as well. They prefer the guideline to be printed in bigger size. However, the A4 size guideline is only printed for evaluation purpose. Since the guideline comprises of textual elements, Expert 1 also suggested the use of application icon which is more attractive.

Second phase of evaluation involved validation of the proposed guideline with the intention to check whether the proposed guideline meets the requirements and expectations. For validating purpose, prototyping method is used. A prototype of dyslexic game is developed, namely DysRedia by taking considerations the criteria listed in the proposed guideline. DysRedia is a proof of concept of the proposed guideline. Fig. 11 shows selected interfaces of DysRedia prototype.



Fig. 11. Selected Interfaces of DysRedia.

Version 1.0 of DysRedia has been demonstrated to two teachers and students at Dyslexia Incubator, School of Computing. 34 children had experienced playing and testing it. DysRedia has been improved by taking considerations of their responses and feedback. For example, letters were initially in upper case have been changed to lower case as shown in Fig. 12.



Fig. 12. The Improved Version of DysRedia.

## V. CONCLUSION

A design guideline for Dyslexic Intervention Games has been successfully designed and evaluated. A significant contribution of this study is the criteria and guidelines for Dyslexic game which will benefit game developers, practitioners, and educationist who are directly involved with Dyslexic children. The proposed guideline can serve in assisting them in designing and developing game applications for dyslexic children.

This study also presented DysRedia, a game which is designed and developed based on the proposed design guideline as a proof of concept to the proposed design guideline. With high acceptance of the game, it is supported that the proposed design guideline is an appropriate design that follow gamification concept, friendly to target audience, and Dyslexic acceptance. The proposed criteria and guidelines can be adapted to other similar domains, such as special needs education therapy, particularly involving with Dyslexic children.

This study is aligned and significant to Sustainable Development Goals (SDG) three and four, Good Health and Well-being and Quality Education respectively. The outcome of this study could contribute in improving the reading ability among dyslexic children. Future works might consider different type of evaluation of the proposed guideline involving bigger audience of testers.

## ACKNOWLEDGMENT

This research is funded by Universiti Utara Malaysia (UUM) through the University Grant [SO code: 14590]. The authors fully acknowledged UUM for the approved fund, which makes this important research viable and effective. Credit also goes to our game developer, Robin Chan.

## REFERENCES

- [1] G. Schulte-Körne, "The prevention, diagnosis, and treatment of dyslexia," *Dtsch. Arztebl. Int.*, vol. 107, no. 41, p. 718, 2010.
- [2] E. Ferrer, B. A. Shaywitz, J. M. Holahan, K. Marchione, and S. E. Shaywitz, "Uncoupling of reading and IQ over time: Empirical evidence for a definition of dyslexia," *Psychol. Sci.*, vol. 21, no. 1, pp. 93–101, 2010.
- [3] S. O. Wajuihian and K. S. Naidoo, "Dyslexia: An overview," *African Vis. Eye Heal.*, vol. 70, no. 2, pp. 89–98, 2011.
- [4] International Dyslexia Association, "Dyslexia basics," 2020. [Online]. Available: <https://dyslexiaida.org/dyslexia-basics-2/>. [Accessed: 17-Jun-2022].
- [5] Special Education Division Ministry of Education, "Special Education Data 2014," *Minist. Educ.*, 2014.
- [6] Socio-economic & Environmental Research Institute Penang, "Learning difficulties among Standard 1 pupils in Penang. Penang: Socio-economic & Environmental Research Institute Penang," 2003.
- [7] H. Hussin, "Mobile Dyslexic Specialized Digital Game-based Learning Object for Learning Letters (DOLL)," 2012.
- [8] J. Glazzard, "The impact of dyslexia on pupils' self-esteem," *Support Learn.*, vol. 25, no. 2, pp. 63–69, 2010.
- [9] D. F. Horrobin, A. I. M. Glen, and C. J. Hudson, "Possible relevance of phospholipid abnormalities and genetic interactions in psychiatric disorders: the relationship between dyslexia and schizophrenia," *Med. Hypotheses*, vol. 45, no. 6, pp. 605–613, 1995.
- [10] A. J. Richardson, "Dyslexia, Dyspraxia and SDHC-Can Nutrition Help?," *Oxford Dyslexia Res. Trust*, pp. 1–10, 2002.

- [11] H. L. Swanson, *Interventions for students with learning disabilities: A meta-analysis of treatment outcomes*. Guilford Press, 1999.
- [12] S. Z. Ahmad, N. N. A. N. Ludin, H. M. Ekhsan, A. F. Rosmani, and M. H. Ismail, "Bijak Membaca—Applying Phonic Reading Technique and Multisensory Approach with interactive multimedia for dyslexia children," in 2012 IEEE Colloquium on Humanities, Science and Engineering (CHUSER), 2012, pp. 554–559.
- [13] T. Cuschieri, R. Khaled, V. E. Farrugia, H. P. Martinez, and G. N. Yannakakis, "The iLearnRW game: support for students with Dyslexia in class and at home," in 2014 6th international conference on games and virtual worlds for serious applications (VS-GAMES), 2014, pp. 1–2.
- [14] P. A. Di Tore, S. Di Tore, L. A. Ludovico, and G. R. Mangione, "Madrigale: a multimedia application for dyslexia and reading improvement gamifying learning experience," in 2014 International Conference on Intelligent Networking and Collaborative Systems, 2014, pp. 486–491.
- [15] L. Rello, C. Bayarri, and A. Gorriz, "Dyslexia exercises on my tablet are more fun," in Proceedings of the 10th International Cross-Disciplinary Conference on Web Accessibility, 2013, pp. 1–2.
- [16] M. 'Azizi C. Sulaiman and A. Ban, "User interface guidelines for dyslexic game-based learning on selected usability test method," *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 8, no. 1.4 S1, pp. 439–445, 2019, doi: 10.30534/ijatcse/2019/6981.42019.
- [17] L. Van den Audenaeren et al., "DYSL-X: Design of a tablet game for early risk detection of dyslexia in preschoolers," in Games for health, Springer, 2013, pp. 257–266.
- [18] H. Husni, Z. Jamaludin, and F. A. Aziz, "Dyslexic Children's Reading Application: Design For Affection," *J. Inf. Commun. Technol.*, vol. 12, pp. 1–19, 2013.
- [19] M. Habib, "The neurological basis of developmental dyslexia: an overview and working hypothesis," *Brain*, vol. 123, no. 12, pp. 2373–2399, 2000.
- [20] A. W. Alexander and A.-M. Slinger-Constant, "Current status of treatments for dyslexia: Critical review," *J. Child Neurol.*, vol. 19, no. 10, pp. 744–758, 2004.
- [21] G. Reid, *Dyslexia: A practitioner's handbook*. John Wiley & Sons, 2016.
- [22] L. Bazen, M. van den Boer, P. F. de Jong, and E. H. de Bree, "Early and late diagnosed dyslexia in secondary school: Performance on literacy skills and cognitive correlates," *Dyslexia*, vol. 26, no. 4, pp. 359–376, 2020.
- [23] W. M. R. W. Mohammad, "Dyslexia in the aspect of Malay language spelling," *Int. J. Acad. Res. Bus. Soc. Sci.*, vol. 2, no. 1, p. 308, 2012.
- [24] R. A. Bolhasan, "A study of dyslexia among primary school students in Sarawak, Malaysia," *Sch. Dr. Stud. (European Union) J.*, vol. 1, no. 1, pp. 250–268, 2009.
- [25] A. J. Fawcett and R. I. Nicolson, "Naming speed in children with dyslexia," *J. Learn. Disabil.*, vol. 27, no. 10, pp. 641–646, 1994.
- [26] C. Jamieson and S. Simpson, "Spelling: Challenges and strategies for the dyslexic learner and the teacher," Whurr, 2006.
- [27] M. Snowling, "Dyslexia as a phonological deficit: Evidence and implications," *Child Psychol. Psychiatry Rev.*, vol. 3, no. 1, pp. 4–11, 1998.
- [28] L. C. Ehri, "Reading processes, acquisition, and instructional implications," *Dyslexia Lit. Theory Pract.*, vol. 167, p. 186, 2002.
- [29] N. Goulrandis, "Assessing reading and spelling skills," *Dyslexia speech Lang.*, pp. 98–127, 2006.
- [30] L. W. Lee and K. Wheldall, "Acquisition of Malay word recognition skills: lessons from low-progress early readers," *Dyslexia*, vol. 17, no. 1, pp. 19–37, 2011.
- [31] N. A. M. Yuzaidey, N. C. Din, M. Ahmad, N. Ibrahim, R. A. Razak, and D. Harun, "Interventions for children with dyslexia: A review on current intervention methods," *Med J Malaysia*, vol. 73, no. 5, p. 311, 2018.
- [32] E. Arnbak and C. Elbro, "The effects of morphological awareness training on the reading and spelling skills of young dyslexics," *Scand. J. Educ. Res.*, vol. 44, no. 3, pp. 229–251, 2000.
- [33] M. Kast, M. Meyer, C. Vögeli, M. Gross, and L. Jäncke, "Computer-based multisensory learning in children with developmental dyslexia," *Restor. Neurol. Neurosci.*, vol. 25, no. 3–4, pp. 355–369, 2007.
- [34] L. L. Wah, "The Davis model of dyslexia intervention: Lessons from one child," *Editor. Board*, p. 133, 2010.
- [35] Y. Luo, J. Wang, H. Wu, D. Zhu, and Y. Zhang, "Working-memory training improves developmental dyslexia in Chinese children," *Neural Regen. Res.*, vol. 8, no. 5, p. 452, 2013.
- [36] S. Franceschini, S. Bertoni, L. Ronconi, M. Molteni, S. Gori, and A. Facoetti, "'Shall we play a game?': Improving reading through action video games in developmental dyslexia," *Curr. Dev. Disord. reports*, vol. 2, no. 4, pp. 318–329, 2015.
- [37] Y. Qian and H.-Y. Bi, "The effect of magnocellular-based visual-motor intervention on Chinese children with developmental dyslexia," *Front. Psychol.*, vol. 6, p. 1529, 2015.
- [38] S. Nourbakhsh, M. Mansor, M. Baba, and Z. Madon, "The effects of multisensory method and cognitive skills training on perceptual performance and reading ability among dyslexic students in Tehran-Iran," *Int. J. Psychol. Stud.*, vol. 5, no. 2, pp. 92–99, 2013.
- [39] R. M. Majzub, M. A. Abdullah, and Z. Aziz, "Effects of a multisensory programme on dyslexic students: Identification and mastery of the alphabet," *Res. J. Appl. Sci.*, vol. 7, no. 7, pp. 340–343, 2012.
- [40] V. Subramaniam, V. K. Mallan, and N. H. C. Mat, "Multi-senses explication activities module for dyslexic children in Malaysia," *Asian Soc. Sci.*, vol. 9, no. 7, p. 241, 2013.
- [41] J. Ohene-Djan and R. Begum, "Multisensory games for dyslexic children," in 2008 Eighth IEEE International Conference on Advanced Learning Technologies, 2008, pp. 1040–1041.
- [42] C. S.-H. Ho, E. Y.-C. Lam, and A. Au, "The effectiveness of multisensory training in improving reading and writing skills of Chinese dyslexic children," *Psychologia*, vol. 44, no. 4, pp. 269–280, 2001.
- [43] S. Franceschini, S. Gori, M. Ruffino, S. Viola, M. Molteni, and A. Facoetti, "Action video games make dyslexic children read better," *Curr. Biol.*, vol. 23, no. 6, pp. 462–466, 2013.
- [44] N. Fusco, G. D. Germano, and S. A. Capellini, "Efficacy of a perceptual and visual-motor skill intervention program for students with dyslexia," in *CoDAS*, 2015, vol. 27, pp. 128–134.
- [45] P. Tzouveli, A. Schmidt, M. Schneider, A. Symvonis, and S. Kollias, "Adaptive reading assistance for the inclusion of students with dyslexia: The AGENT-DYSL approach," in 2008 Eighth IEEE International Conference on Advanced Learning Technologies, 2008, pp. 167–171.
- [46] D. Coyle, M. Matthews, J. Sharry, A. Nisbet, and G. Doherty, "Personal Investigator: A therapeutic 3D game for adolescent psychotherapy," *Interact. Technol. smart Educ.*, 2005.
- [47] J. Li, Y.-L. Theng, and S. Foo, "Game-based digital interventions for depression therapy: a systematic review and meta-analysis," *Cyberpsychology, Behav. Soc. Netw.*, vol. 17, no. 8, pp. 519–527, 2014.
- [48] J. Cheng, C. Putnam, and D. C. Rusch, "Towards efficacy-centered game design patterns for brain injury rehabilitation: A data-driven approach," in Proceedings of the 17th International ACM SIGACCESS Conference on Computers & Accessibility, 2015, pp. 291–299.
- [49] Y. Li, W. Fontijn, and P. Markopoulos, "A tangible tabletop game supporting therapy of children with cerebral palsy," in International Conference on Fun and Games, 2008, pp. 182–193.
- [50] S. K. Tatla et al., "Therapists' perceptions of social media and video game technologies in upper limb rehabilitation," *JMIR serious games*, vol. 3, no. 1, p. e3401, 2015.
- [51] N. A. Bartolomé, A. M. Zorrilla, and B. G. Zapirain, "Dyslexia diagnosis in reading stage through the use of games at school," in 2012 17th International Conference on Computer Games (CGAMES), 2012, pp. 12–17.
- [52] L. Rello, S. Subirats, and J. P. Bigham, "An online chess game designed for people with dyslexia," in Proceedings of the 13th International Web for All Conference, 2016, pp. 1–8.
- [53] O. Gaggi et al., "Serious games for early identification of developmental dyslexia," *Comput. Entertain.*, vol. 15, no. 2, pp. 1–24, 2017.

- [54] M. H. L. Abdullah, S. Hisham, and S. Parumo, "MyLexics: an assistive courseware for Dyslexic children to learn basic Malay language," ACM SIGACCESS Access. Comput., no. 95, pp. 3–9, 2009.
- [55] L. Rello, C. Bayarri, and A. Gorriz, "What is wrong with this word? Dysegxia: a game for children with dyslexia," in Proceedings of the 14th international ACM SIGACCESS conference on Computers and accessibility, 2012, pp. 219–220.
- [56] M. R. U. Saputra and M. Risqi, "LexiPal: Design, implementation and evaluation of gamification on learning application for dyslexia," Int. J. Comput. Appl., vol. 131, no. 7, pp. 37–43, 2015.
- [57] H. Holz et al., "Prosodiya—a mobile game for german dyslexic children," in International conference on games and learning alliance, 2017, pp. 73–82.
- [58] M. Ronimus, K. Eklund, L. Pesu, and H. Lyytinen, "Supporting struggling readers with digital game-based learning," Educ. Technol. Res. Dev., vol. 67, no. 3, pp. 639–663, 2019.
- [59] D. Gooch, A. Vasalou, L. Benton, and R. Khaled, "Using gamification to motivate students with dyslexia," in Proceedings of the 2016 CHI Conference on human factors in computing systems, 2016, pp. 969–980.
- [60] C. Singleton and F. Simmons, "An evaluation of Wordshark in the classroom," Br. J. Educ. Technol., vol. 32, no. 3, pp. 317–330, 2001.
- [61] R. T. Bigueras, M. C. A. Arispe, J. O. Torio, and D. E. Maligat Jr, "Mobile Game-Based Learning to Enhance the Reading Performance of Dyslexic Children," Int. J., vol. 9, no. 1.3, 2020.
- [62] S. Franceschini and S. Bertoni, "Improving action video games abilities increases the phonological decoding speed and phonological short-term memory in children with developmental dyslexia," Neuropsychologia, vol. 130, pp. 100–106, 2019.
- [63] R. Görgen, S. Huemer, G. Schulte-Körne, and K. Moll, "Evaluation of a digital game-based reading training for German children with reading disorder," Comput. Educ., vol. 150, no. January, 2020, doi: 10.1016/j.compedu.2020.103834.
- [64] H. Holz, B. Beuttler, and M. Ninaus, "Design rationales of a mobile game-based intervention for german dyslexic children," in Proceedings of the 2018 annual symposium on computer-human interaction in play companion extended abstracts, 2018, pp. 205–219.
- [65] K. Preffers, "A Design Science Research Methodology for information System," J. Manag. Inf. Syst., vol. 24, no. 3, pp. 45–78, 2007.
- [66] D. Moher, A. Liberati, J. Tetzlaff, D. G. Altman, and P. Group\*, "Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement," Ann. Intern. Med., vol. 151, no. 4, pp. 264–269, 2009.

# The Effectiveness of Gamification for Students' Engagement in Technical and Vocational Education and Training

Laily Abu Samah, Amirah Ismail, Mohammad Kamrul Hasan  
Faculty of Information Science and Technology  
Universiti Kebangsaan Malaysia (UKM)  
Bangi, Selangor, Malaysia

**Abstract**—The transformation of Technical and Vocational Education and Training (TVET) prioritizes by the national education convention to meet the needs of the industry through improving student skills and the quality of related systems. One of the transformations is practicing blended learning, such as a flipped classroom, to produce better quality student learning outcomes. However, based on previous studies, there are difficulties in maintaining student engagement during learning activities, even though blended learning offers some advantages. Therefore, this study suggests the development of a mobile application using gamification as a solution to enhance student participation. This paper proposes the design and development research (DDR) approach with the adaptation of the ADDIE model to build a learning content prototype. It involves five phases: analysis, design, development, implementation, and evaluation. The study participants consisted of two groups of students in the 1st semester of the Interactive Multimedia course from two different TVET institutions who were cleft into a control group and an experimental group. The experimental group is gamified, whereas the control group is not. The study evaluation uses two instruments: a test to compare students' understanding of both groups and an activity log to track the experimental group's use of the prototype. According to the findings, gamification during learning activities can increase student engagement by boosting performance through a more significant pre-and post-test mean score difference and creating a positive learning experience. Additionally, mobile applications with the gamification concept can be employed extensively in various TVET courses to encourage student learning performance.

**Keywords**—Technical and vocational education and training (TVET); flipped classroom; engagement; gamification; mobile application

## I. INTRODUCTION

UNESCO defines Technical and Vocational Education and Training (TVET) as a term that refers to aspects of education. It involves alternatives to academic education, the study of technology and related sciences, and an environment to acquire and apply knowledge, skills, and attitudes related to employment in various sectors of the economy and social life. The measurement of TVET students' competence refers to job analysis developed through the coordination of industry experts, skilled workers, and teaching experts by field to ensure

that the implementation of TVET meets and is in line with the needs of the industry [1].

Various approaches are used during learning activities to guarantee that students attain the essential competencies, one of which is blended learning. Blended learning is a blend of face-to-face or online learning activities that occur inside or outside of the classroom, such as conversations in group work, hands-on practice, presentations, and project-based solutions [2]. It was detected to be more successful at enhancing student engagement.

Although blended learning has a positive effect on student engagement, according to [3], there are difficulties in maintaining it. To sustain engagement potential and achieve learning objectives, students must be wise in how well they manage their attitude and autonomy during learning activities.

According to previous studies, gamification has been widely used in TVET, albeit it is unclear whether this manages to boost engagement [4]. Therefore, this study suggests developing a mobile application using gamification as a solution to assess student engagement by their achievement and learning experiences. Since students spend most of their time on their phones, mobile applications have emerged as the best way to motivate them to learn.

The implementation of this study is vital in contributing towards:

- Make students engage in learning activities through gamification to increase student performance and reduce the dropout rate.
- It adds to research on the wide use of gamification in the TVET environment but is less prominent.
- It can be extended to the entire TVET, whether public or private TVET institutions, in various areas of TVET.

## II. LITERATURE REVIEW

This section explains the synthesis of information obtained to assess student engagement in TVET blended learning through gamification. It is divided into several parts, starting with the introduction to TVET, the flipped classroom practiced in TVET institutions, the potential of student engagement in



learning activities, and the use of gamification elements to develop mobile application prototypes for self-learning.

#### A. Technical and Vocational Education and Training (TVET)

TVET plays a role in producing a skilled workforce in various fields through training that allows students to acquire knowledge and skills. It shapes students to have a lifelong learning mindset and be capable as employers who create jobs. TVET also provides individuals with expertise and skills appropriate to the job market to address the global unemployment problem that will produce competent and creative workers who function as agents of sustainability in the workplace [5].

Several aspects need to consider for effective TVET implementation listed as:

- The use and influence of technology in learning activities, like the importance of ICT to solve problems creatively and analytically through various applications, software, and devices [5].
- Instructors' preparation ensures that learning activities efficiently run where they need to master the knowledge and skills in the field by being able to explain and demonstrate correct and safe work steps and answer any questions from students [6].
- TVET-related systems are understood and implemented by all parties, for example, using job analysis correctly as a reference to carry out the learning process along with the right equipment, work steps, and technical information arranged according to the difficulty level [1].

Aspects mentioned, such as the influence of ICT, teacher preparation, and the system set, encourage the diversity of the implementation of learning activities in the TVET environment. One method that has gained attention is blended learning, which meets TVET learning activities by emphasizing specific jobs' theoretical and practical components. The following section will discuss blended learning.

#### B. Flipped Classroom

The learning environment needs to provide space for communication, collaboration, creativity, and critical thinking in making decisions, developing strategies, and solving problems with the help of technology [7]. This conducive condition is necessary to avoid boredom that limits the ability to perform tasks and creates a feeling of lack of interest, loss of motivation, and absence of student engagement [8].

Therefore, TVET practices blended learning to create a conducive learning environment. A famous example of blended learning is the flipped classroom. Students prepare beforehand using learning materials such as presentation slides or videos before undergoing face-to-face learning with instructor monitoring through various activities such as discussions, presentations, drills, group assignments, and assessments [8].

Comparison between flipped and traditional classrooms is the better way to understand it. Implementing a conventional

classroom is through the delivery of learning content by the instructor in the class at a set time and period. Then students must complete the tasks provided after the end of the study [9]. Meanwhile, flipped classroom implementation contradicts traditional classrooms, as shown in Fig. 1.

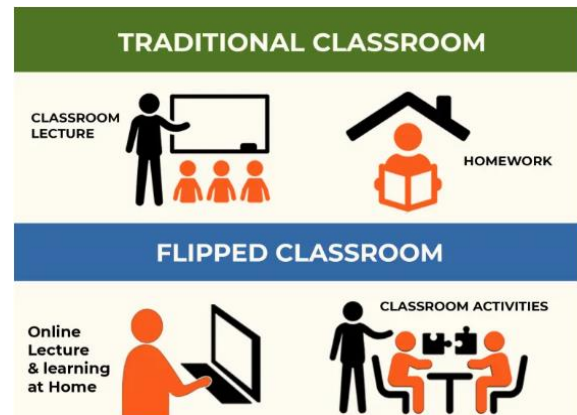


Fig. 1. Traditional Versus Flipped Classroom (Source: <https://edtechimpact.com/news/flipping-the-classroom-ultimate-guide>).

Although blended learning practices in the TVET environment positively impact students, engagement is challenging to maintain. This hardship pushes by several factors, such as the diversity of student characters, less effective learning materials, and constraints in the use of learning technology. The next section of the literature answers several points related to student engagement in learning activities.

#### C. Engagement

The foundation of high-quality learning is engagement, defined as the use of time and energy to carry out an action or task differently impacted by many circumstances [10]. Students' good attendance, commitment, interaction in learning activities [11], and valuable personalities, including satisfaction, success, belonging, enjoyment, liking, skills, competence, perseverance, motivation, and courage, are examples of engagement [4] [12].

Student engagement in learning activities is measured using numerous instruments related to various items. Some use the National Survey of Student Engagement (NSSE), K-12 classroom engagement scale, Student Engagement Questionnaire (SEQ), and a combination of positive and negative affect schedule and presence questionnaires [13]. However, most of the instruments employed focus on the three main measurement criteria' behavioral, cognitive, and emotional.

Behavioral criteria were discovered by looking at how engaged and diligent students are in their studies and how willing they are to ask for assistance when necessary [14]. These criteria also affect utilizing qualitative elements that require effort, attention, and persistence while being observed [4].

Cognitive criteria implicate the amount of effort and time required to comprehend the work, the drive to overcome shortcomings, the ease with which one can adjust to problem-

solving, and the achievement of learning success [14]. For easier understanding, this criterion requires efforts of academic knowledge's intellectual components [15].

Emotional criteria are determined by looking at positive emotions, such as excitement, joy, and confidence, as well as negative ones, such as boredom, frustration, and anxiety [13]. It also measures expression through feelings or reactions between a combination of physiological and psychomotor components positively or negatively, including pride and anger [4].

When implementing learning activities, it can be troublesome to keep students' attention because they come from different backgrounds and have different learning styles. Hence, it is necessary to set difficulty levels starting with easy, medium, and challenging levels throughout learning activities [16].

How can engagement be maintained to ensure learning activities achieve the set learning outcomes? This question is always floating around, and the implementation of various methods to address this issue. One way is using gamification in learning activities described in detail in the next section.

#### D. Gamification

Instructors are continually experimenting with a new pedagogical method to capture students' attention, motivate them, and engage them in learning activities. Thus, rather than traditional learning methods, digital computer games are customized to create an enjoyable and engaging learning environment for students [17]. The adaptation of a game into a non-game condition is key to gamification. As a result, the definition of gamification is the use of game elements or mechanics in non-game contexts [18].

Explicit knowledge is required when comparing gamification in learning activities to other methods that also use the basics of games, such as game-based learning (GBL), serious games, and simulations. GBL employs the power of games to engage students in learning activities [19]. On the other hand, serious games resemble gaming design worlds that solve problems unrelated to enjoyment [20]. While the simulation parallels a serious game, the main objective is training in the military, medicine, and aviation fields [17].

Gamification includes a variety of qualities that help it achieve its purpose. It provides rewards and develops motivation [21] to complete specific tasks. It makes learning material more dynamic, innovative, and appealing, encouraging participation and boosting understanding of the learning substances [22]. It promotes various active and successful learning strategies by maintaining attention and interest in all learning tasks [23]. It can also be an alternate strategy that adds value to normal learning activities by giving a more engaging experience through gamification elements [14]. Furthermore, gamification persuades students to complete assignments despite exhaustion [17].

Gamification aims to lower dropout rates by providing students with practical learning methods and creating a pleasant learning environment [22]. It can increase students' enjoyment and engagement to kindle their interest in studying

and obtain better results [21]. It can motivate and encourage student competitiveness by improving student happiness, effectiveness, and efficiency in learning activities [14]. When students face challenging topics and limited time, it eliminates challenges and solves problems during learning activities by integrating learning activities appropriately [18].

Application developers and instructors must work together to ensure that the gamification design delivers maximum benefits and effectively enhances student engagement [10]. One of the primary design criteria is incorporating relevant and purposeful gamification elements into the learning content via set objectives. By delivering clear, intuitive, and pleasant learning content, gamification elements should generate a good learning experience [21]. Table I depicts the use of gamification elements related to student engagement based on previous studies.

Besides that, gamification interface design must incorporate seamless navigation [21], an exciting narrative adjustment, and an acceptable combination of text, graphics, colors, and animations [17]. It is to ensure that students do not lose focus due to the excessive amount of gamification design so that they stray from the original learning goal [24]. The processing of learning content into gamification design needs to emphasize learning strategies so that students obtain quality learning results and experiences. Among the techniques practiced is segmentation, which breaks down learning material at a rate students can accept for knowledge retention [25]. The arrangement of learning content also needs to follow levels starting from low, medium, and high levels to meet the needs and abilities of students from various backgrounds [17].

The coming section explains the research methodology based on the problem statement and literature review. This section details the steps implemented to develop a mobile learning application prototype by including gamification elements identified to increase student engagement.

TABLE I. GAMIFICATION ELEMENTS RELATED TO STUDENT ENGAGEMENT BASED ON PREVIOUS STUDIES

Gamification elements	Source											
	[4]	[10]	[11]	[14]	[15]	[17]	[18]	[19]	[21]	[22]	[23]	[24]
Badges		X	X		X		X	X	X			
Challenges		X							X			
Leaderboard	X								X			
Leveling	X	X	X		X		X	X	X	X		X
Points	X	X	X		X		X	X	X	X	X	X
Unlock content									X			
Avatar		X								X		
Progress bar	X		X		X			X		X		
Rewards/Awards				X		X	X			X	X	
Feedback			X	X	X	X						
Time pressure/ limit			X		X			X				X
Life											X	

### III. METHODOLOGY

The study employs experiential learning theory, an essential theory based on cognitive results. This theory states that learning activities repeatedly occur through experience modification, observational reflection, abstract conceptualization, and active experimentation [26]. The repeating process is made feasible by learning exercises over a developed gamified mobile app prototype. The study occupies the design and development research (DDR) approach with the adaptation of the ADDIE model depicted in Fig. 2.

The selection of DDR as a research methodology is because DDR involves a systematic and organized process consisting of three stages: designing, developing, and evaluating the success of mobile learning application prototypes to obtain empirical evidence based on the collection and analysis of data from experiments conducted. The adapted ADDIE model into DDR consists of five phases: analysis, design, development, implementation, and evaluation, which are explained in more detail hereafter.

#### A. Analysis Phase

The analysis phase involves the analysis and setting of some criteria to launch the implementation of the study. It determines study participants, prototype users, learning content, and authoring tools with appropriate gamification elements.

The participants chosen were first-semester students' of Software Technology (Interactive Multimedia) from two different TVET institutions split into a control group and an experimental group. Both groups underwent face-to-face, blended learning activities, while only the experimental group had to use the gamified prototype as an addition to self-learning.

The module as the prototype's content is Image Editing from the five available modules. The topics involved are

shooting, selecting, editing, and saving photos. It is used with the instructor's help to optimize characteristics of knowledge, abilities, and attitudes in the prototype.

Furthermore, the selection of proper authoring tools ensures that the process of producing and testing the prototype achieves the study's goal. The research uses Buildbox software because it offers prototyping features with an excellent 2D graphic resolution display, appropriate gamification elements, and the drag-and-drop concept.

#### B. Design Phase

The prototype consists of four main parts according to sub-topics: notes, quizzes, assessments, and games. The prototype's content should comply with the TVET learning environment, including knowledge, skills, and attitude competencies. As in Table II, prototypes are developed by including eight gamification elements to trigger student engagement.

The prototype also applies Mayer's multimedia design principles to support learning activities to be more quality and effective. It also uses Jakob Nielsen's heuristic evaluation to ensure the display of the prototype works well.

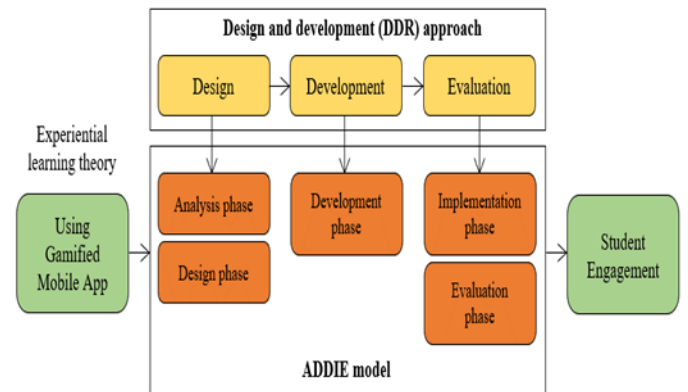


Fig. 2. DDR Approach with the Adaption of the ADDIE Model.

TABLE II. GAMIFICATION ELEMENTS USED IN THE PROTOTYPE

Part	Gamification elements	Function
Notes	Leveling <sub>1</sub>	Before moving on to the subsequent sub-topic, be sure the sub-topic before has been finished.
	Progress bar <sub>2</sub>	Disclose the progress status of the reviewing notes.
	Badges <sub>3</sub>	The current sub-topic was reviewed and determined to be ready for a quiz or progression to the next sub-topic.
Quizzes	Points <sub>4</sub>	Answers earn points. If the answer is correct, points count; if incorrect, no points count.
	Feedback <sub>5</sub>	Feedback is provided based on answers. If the answer is correct, proceed to the next question. The necessary notes are displayed if the answer is incorrect and returned to the current question.
	Life <sub>6</sub>	Life is permitted to provide answers. If the answer is incorrect, life deducts, and students repeat the quiz if there is no more life.
	Leveling <sub>1</sub>	Students must earn all possible points to be qualified to respond to the assessment question or go on to the following subtopic.
Assessments	Time limit <sub>7</sub>	Allow time to respond to questions. No points if the timer ran out and skipped the question.
	Points <sub>4</sub>	Answers earn points. If the answer is correct, points count; if incorrect, no points count.
	Leveling <sub>1</sub>	Must complete specific parts (notes and quizzes) to be qualified to respond to the following assessment question.
Games	Challenges <sub>8</sub>	Make the game challenging by requiring players to collect and avoid specific things.
	Life <sub>6</sub>	Set up a gaming environment. Failure to dodge obstacles results in reduced life, and when life is gone, the game is over.
	Points <sub>4</sub>	Make a competition within the game to increase current points for the next round.

<sup>a</sup>. Total of gamification used: 1,2,3,4,5,6,7,8



Fig. 3. Some Prototype Interfaces that Incorporate Gamification Elements.

The quizzes and assessments structure meets the criteria of difficulty level and skill type formed through the Test Specification Table. The difficulty level refers to the value of the question hardship starting from low, medium, and high based on the student's ability to answer the question using the ratio 1 (low): 2 (medium): 1 (high). In contrast, the skill type refers to a category built on three aspects, consisting of theoretical, procedural, and attitude parts, using the ratio 6 (theory): 3 (procedure): 1 (attitude).

The game intends to give students a space to rest for a while from entirely focusing on learning activities through other parts of this prototype. However, the game produced revolves around graphics related to the learning material.

### C. Development Phase

This phase entails forming a prototype and continual testing to guarantee that it is functional. The prototype pre-use is a series of usability tests conducted with students, instructors, and developer experts utilizing the thinking-aloud method. Each examiner made a clear vocal comment while using the prototype, and the researcher recorded the statements.

Visual and functional characteristics that affect prototype performance, such as appropriateness and precision of navigation, clear writing, quality visuals, effective interface display, and efficient learning content, are reviewed. Fig. 3 depicts some prototype interfaces that incorporate gamification elements.

### D. Implementation Phase

The implementation phase includes a pre-test followed by the prototype usage within a specific time frame and ending with a post-test by 23 students in each group. The pre-and post-tests use the 40 same multiple choice questions (MCQ) in different positions to assess student comprehension.

The experimental groups (prototype users) must update the activity logs within two weeks of utilizing the prototype. This step intends to reduce the possibility of students becoming disinterested in self-learning. Meanwhile, the control group only underwent blended learning activities with the instructors.

The activity logs provided the student's prototype progress throughout the self-study session. Students only answer the items provided through yes or no options, record the date for each stage, and answer a few short questions. Intending to make it easier for students to provide the necessary information, gain initial exposure related to the prototype content, and allows students to focus while using this prototype.

### E. Evaluation Phase

The evaluation phase is the study's final stage to determine gamification's effectiveness on student engagement in TVET blended learning. Evaluation of the construction of a hypothesis is conducted based on pre-and post-test scores as follows:

- $h_0$  – no significant difference between the pre-test and post-test scores of the control group
- $h_1$  – no significant difference between the pre-test and post-test scores of the experimental group

In addition to the hypothesis findings, the study also analyzed the activity log updated by the experimental group.

## IV. RESULTS

### A. Pre-and Post-Test Scores

A quantitative analysis was conducted on the pre-and post-test scores by a total of 46 students from both control and experimental groups using SPSS software. Before the execution of the investigation on the constructed hypothesis, a normality test runs on the entire score obtained to determine the normal distribution of the scores. Table III shows the Shapiro-Wilk normality test results for the control and experimental groups' pre-and post-tests.

TABLE III. SHAPIRO-WILK NORMALITY TEST RESULTS

Group	Test	Statistic	df	Sig
Control	Pre	0.954	23	<b>0.545</b>
	Post	0.956	23	<b>0.389</b>
Experimental	Pre	0.923	23	<b>0.079</b>
	Post	0.953	23	<b>0.342</b>

Based on the table shown that the Shapiro-Wilk value for the entire score is determined as normally distributed with a significant rate of  $p > 0.05$ . As a result, parametric tests can be performed based on the overall score.

A paired sample t-test was conducted on pre-and post-test scores to prove the hypothesis built based on a significant value of  $p < 0.05$  as follows:

- " $h_0$  – no significant difference between the pre-test and post-test scores of the control group". Value  $p = 0.001$ , then  $h_0$  – rejected.
- " $h_1$  – no significant difference between the pre-test and post-test scores of the experimental group". Value  $p = 0.001$ , then  $h_1$  – rejected.

Depending on the paired sample t-test, hypotheses were all rejected because significant differences between the tested variables showed that student understanding increased between the pre-and post-test. However, the improvement achieved by students from the experimental group using gamified application prototypes is more remarkable through a mean difference of 17.52, referring to Table IV. This difference demonstrates that gamification affects student engagement in self-learning by leading to higher learning outcomes.

TABLE IV. THE PRE-AND POST-TEST MEAN DIFFERENCE

Group	Test	Mean Score	Mean score difference
Control	Pre	20.09	9.95
	Post	30.04	
Experimental	Pre	17.83	17.52
	Post	35.35	

### B. Gamified Activity Logs Analysis

An analysis enforced three criteria to analyze the activity logs amended by prototype users. It includes cognitive, behavioral, and emotional factors related to gamification elements influencing student engagement during learning activities. Table V summarizes the activity log's findings through gamification elements.

The gamification design is also vital in boosting the quality of learning activities. Table VI summarizes the engagement measurement of different gamification designs used in the prototype.

TABLE V. THE ACTIVITY LOG'S FINDINGS THROUGH GAMIFICATION ELEMENTS

Assessment criteria	Assessment item	Worksheet item	Finding	Result
Cognitive	Learning repetition	How many attempts to earn full marks when answering the quiz to be eligible to answer the assessment?	The mean of quiz repetitions is 3.32 times.	There is engagement through the retention of knowledge due to the repetition of learning activities.
		Do you read finished notes repeatedly?	100% answered Yes.	
		Did you repeat the completed assessment to improve the score obtained?	100% answered Yes.	
	Assessment score	The obtained score while answering the assessment.	The assessment mean of scores is 7.85 compared to 10 questions for each sub-topic.	There is engagement through good scores while undergoing assessment.
Behavioral	Duration	Duration to collect the badges (complete review of each sub-topic).	The duration mean is 5.47 days compared to 14 days to use the prototype.	There is engagement through attention and persistence due to using the prototype in a short period.
	Motivation	Did level openings by completed sub-topics motivate you to finish the study?	100% answered Yes.	There is engagement through motivation due to the completed sub-topic.
	Focus	Does the length of time to answer questions make you more focused on answering?	100% answered Yes.	There is engagement through the focus given when answering the assessment.
Emotional	Fun	Does earning badges give you joy?	100% answered Yes.	Engagement through emotions shows fun, stress, enthusiasm, and satisfaction during learning activities using certain gamification elements.
	Pressure	Does trying to get full marks would make you pressured?	100% answered Yes.	
	Enthusiastic	Does getting full marks and being eligible to answer the assessment make you enthusiastic?	100% answered Yes.	
	Satisfaction	Are you satisfied with the use of this prototype?	100% answered Yes.	



TABLE VI. ENGAGEMENT MEASUREMENT THROUGH GAMIFICATION DESIGN

Assessment item	Worksheet item	Finding	Result
Segmentation	Are the notes provided precise and easy to understand?	100% answered Yes.	The appropriateness of gamification design is critical. It ensures optimal gain of the positive effects of gamification.
Feedback	Does feedback on wrong answers help you?	100% answered Yes.	
Educational games	Did you learn something even while playing?	100% answered Yes.	A more flexible gamification design is needed so students can take a break from the relatively dense and heavy learning content.
	Do you play while studying?	100% answered Yes.	

## V. DISCUSSIONS

Based on the analysis enforced on two measurement instruments, gamification in TVET blended learning has proven to enhance student engagement during self-learning sessions. Student engagement during learning activities positively impacts students' learning experience, especially the emotions of fun that produce an effective learning environment. Subsequently, learning results increase through better student achievement than activities without gamification.

The main contribution of this study is to develop a mobile learning application prototype into TVET blended learning with gamification to increase the potential of involvement. The developed application prototype can be used as a reference and modified according to the suitability of learning in other TVET fields that emphasize competence from the aspects of knowledge, skills, and attitudes. Moreover, this prototype expects to help facilitate learning and provide additional reference resources for instructors and students.

Several previous studies also support the findings of this study. Gamification provides students a pleasant learning experience by encouraging comprehension, pleasure, and higher concentration. It promotes learning, lowers boredom, and enhances engagement, resulting in competitiveness and improved performance [27]. Students feel satisfied, have better interactions, stress and worry about the evaluation are lessened, and the generation's psychological requirements are met [28]. It boosts students' skills in discovering and solving complicated problems through simplified learning [8]. It increases student motivation as a stimulus for active participation in higher and continual learning performance. As a result, the dropout rate is reduced, particularly in the TVET context [29].

Even though gamification has a beneficial influence, several concerns must be addressed, particularly regarding long-term usage. For example, reducing students' motivation due to rules that prevent access to the next activity if they have not completed the previous one and notifications for completing incomplete activities or reaching a certain level [30].

## VI. CONCLUSIONS

This study successfully created a gamified mobile app prototype to assess the effectiveness of gamification on student engagement in TVET blended learning. Gamification assists students in improving their achievement and having a better learning experience. The prototype can be used in related domains and extended to additional TVET fields as gamified learning resources to ease the current learning process.

This study encountered several limits, including a lack of research on gamification in the context of TVET blended learning environments compared to academic-based education. The limitation slightly disrupts the study flow to gathering the best and most useful reference materials. Next, mobile devices impact the delivery of learning materials because of the limited display size, which causes misunderstanding of simple statements by students, low-quality graphics due to small size, and the difficulty in maintaining uniformity, such as the size of texts and answer selection buttons.

In addition, the application prototype is limited to Android device users only. Application prototype development for other platforms such as IOS devices and websites needs different software or system settings. This process requires allocating a lengthened period and appropriate expertise to enable the prototype on various platforms.

Gamified prototypes have the potential to be expanded as an alternate technique for executing learning activities to enhance student engagement in TVET and other educational domains. Collaboration among diverse stakeholders such as instructors, universities, industries, and application developers is vital in maximizing gamification's benefits. As a result, future study recommendations presents as follows:

- The sample size of the students from the control and experimental groups increased, and the more extended period of the application prototype use to determine a more accurate measurement of potential engagement.
- Further research on more specific gamification elements and designs is needed as a measurement item to identify the existence of student engagement in detail.
- The study is extended through blended learning using gamification for students with disabilities in the TVET environment.

## REFERENCES

- [1] M. A. Dokadawa and M. Ali, "Roles of Occupational Analysis towards Effective Teaching and Learning in Technical Vocational Education and Training (TVET) Institutions," *East African Sch. J Edu Humanit Lit*, vol. 4, no. 5, pp. 215–219, 2021, doi: 10.36349/easjehl.2021.v04i05.002.
- [2] A. Gerber and S. Eybers, "Converting to inclusive online flipped classrooms in response to Covid-19 lockdown," *South African J. High. Educ.*, vol. 35, no. 4, pp. 34–57, 2021, doi: 10.20853/35-4-4285.
- [3] M. M. Mohamad, A. Ahmad, M. H. Yee, T. K. Tee, and A. N. Mohd Nasir, "Implementation of Self-Directed Learning in Enhancing Skills Dedicated to the Community College Teaching Staff," in *Journal of Physics: Conference Series*, Mar. 2021, vol. 1793, no. 1, doi: 10.1088/1742-6596/1793/1/012031.
- [4] J. Jayalath and V. Esichaikul, "Gamification to Enhance Motivation and Engagement in Blended eLearning for Technical and Vocational Education and Training," *Technol. Knowl. Learn.*, 2020, doi: 10.1007/s10758-020-09466-2.



- [5] E. O. Ugwoke, T. O. Olinya, H. C. Anorue, and F. S. Abdullahi, "Policies on Digitalization of Instructions: Implication on Teaching and Learning of TVET Programmes," *Vocat. Tech. Educ. J. (VOTEJ)*, vol. 2, no. 2, pp. 2734–2697, 2020.
- [6] A. S. Nabilah, F. A. N. Yunus, R. Mohd. Bekri, and M. Saiful Hadi, "COMPETENCY OF TVET INSTRUCTOR TOWARD TECHNICAL LESSON IN INSTITUT LATIHAN PERINDUSTRIAN (ILP)," *Online J. TVET Pract.*, vol. 3, no. 1, 2018.
- [7] N. Jalinus, U. Verawardina, Krismadinata, R. Azis Nabawi, and Y. Darma, "Developing Blended Learning Model in Vocational Education Based On 21st Century Integrated Learning and Industrial Revolution 4.0," *Turkish J. Comput. Math. Educ.*, vol. 12, no. 8, pp. 1239–1254, 2021.
- [8] T. Abdian, S. Abdollahifar, and L. Mosalanejad, "Implementation of Gamification from blended learning based on the flex model and efficacy of this program on students: an experiences from Iran, An Quasi-experimental Study," pp. 1–23, 2019, doi: <https://doi.org/10.21203/rs.2.14677/v1>.
- [9] A. M. Nortvig, A. K. Petersen, and S. H. Balle, "A literature review of the factors influencing e-learning and blended learning in relation to learning outcome, student satisfaction and engagement," *Electron. J. e-Learning*, vol. 16, no. 1, pp. 46–55, 2018.
- [10] R. S. Alsawaier, "The effect of gamification on motivation and engagement," *Int. J. Inf. Learn. Technol.*, vol. 35, no. 1, pp. 56–79, 2018, doi: 10.1108/IJILT-02-2017-0009.
- [11] F. L. Khaleel, N. S. Ashaari, and T. S. M. T. Wook, "The impact of gamification on students learning engagement," *Int. J. Electr. Comput. Eng.*, vol. 10, no. 5, pp. 4965–4972, 2020, doi: 10.11591/ijece.v10i5.pp4965-4972.
- [12] L. A. Samah and A. Ismail, "Enhance Motivation and Engagement in Blended e-Learning for TVET Using Gamification," *Proc. Int. Conf. Electr. Eng. Informatics*, 2021, doi: 10.1109/ICEEI52609.2021.9611100.
- [13] L. R. Halverson and C. R. Graham, "Learner engagement in blended learning environments: A conceptual framework," *Online Learn. J.*, vol. 23, no. 2, pp. 145–178, 2019, doi: 10.24059/olj.v23i2.1481.
- [14] K. Korkealehto and P. Siklander, "Enhancing engagement, enjoyment and learning experiences through gamification on an English course for health care students," *Semin. J. Media, Technol. Lifelong Learn.*, vol. 14, no. 1, pp. 13–30, 2018, doi: 10.7577/seminar.2579.
- [15] F. L. Khaleel, N. S. Ashaari, T. S. M. T. Wook, and A. Ismail, "Gamification elements for learning applications," *Int. J. Adv. Sci. Eng. Inf. Technol.*, vol. 6, no. 6, pp. 868–874, 2016, doi: 10.18517/ijaseit.6.6.1379.
- [16] T. S. M. T. Wook, I. Y. Zairon, M. Rahmat, H. A. Dahlan, and S. M. Salleh, "Strategi gamifikasi gaya mentoring pembelajaran aktif dalam kalangan pelajar milineal," *J. Teknol. Mklm. dan Multimed. Asia-Pasifik*, vol. 10, no. 1, pp. 141–155, 2021.
- [17] D. A. Alajaji and A. A. Alshwiah, "Effect of combining gamification and a scavenger hunt on pre-service teachers' perceptions and achievement," *J. Inf. Technol. Educ. Res.*, vol. 20, pp. 283–308, 2021, doi: 10.28945/4809.
- [18] F. L. Khaleel, N. S. Ashaari, and T. S. M. T. Wook, "An empirical study on gamification for learning programming language website," *J. Teknol.*, vol. 81, no. 2, pp. 151–162, 2019, doi: 10.11113/jt.v81i1.1133.
- [19] F. L. Khaleel, N. S. Ashaari, T. S. M. T. Wook, and A. Ismail, "Gamification-based learning framework for a programming course," *Proc. 2017 6th Int. Conf. Electr. Eng. Informatics Sustain. Soc. Through Digit. Innov. ICEEI 2017*, vol. 2017-Novem, pp. 1–6, 2018, doi: 10.1109/ICEEI.2017.8312377.
- [20] J. Jayalath and V. Esichaikul, "Gamification-embedded eLearning courses for the learner success of competency based education : Case of Technical and Vocational Education and Training," *8th Pan-Commonwealth Forum Open Learn.*, no. November, 2016.
- [21] A. Hansch, C. Newman, and T. Schildhauer, "Fostering Engagement with Gamification: Review of Current Practices on Online Learning Platforms," *SSRN Electron. J.*, 2015, doi: 10.2139/ssrn.2694736.
- [22] S. N. M. Mohamad, N. S. S. Sazali, and M. A. Mohd Salleh, "Gamification Approach in Education to Increase Learning Engagement," *Int. J. Humanit. Arts Soc. Sci.*, vol. 4, no. 1, pp. 22–32, 2018, doi: 10.20469/ijhss.4.10003-1.
- [23] S. N. M. Mohamad, M. A. M. Salleh, M. Hakim, A. Hamid, L. K. M. Sui, and C. K. N. C. K. Mohd, "Adaptive Learning Strategies with Gamification to Enhance Learning Engagement," *Indian J. Sci. Technol.*, vol. 12, no. 31, 2019, doi: 10.17485/ijst/2019/v12i31/146871.
- [24] S. A. Menon, "Designing Online Materials for Blended Learning: Optimising on BookWidgets," *Int. J. Linguist. Lit. Transl.*, vol. 2, no. 3, pp. 166–174, 2019, doi: 10.32996/ijllt.2019.2.3.19.
- [25] M. Tan and K. F. Hew, "Incorporating meaningful gamification in a blended learning research methods class: Examining student learning, engagement, and affective outcomes," *Australas. J. Educ. Technol.*, vol. 32, no. 5, pp. 19–34, 2016, doi: 10.14742/ajet.2232.
- [26] A. Y. Kolb and D. A. Kolb, "Experiential learning theory: A dynamic, holistic approach to management learning, education and development," *Armstrong Manag. Learn. Edu. Dev.*, pp. 42–68, 2011, doi: 10.4135/9780857021038.n3.
- [27] A. Henukh and Y. Guntara, "Analyzing the response of learners to use kahoot as gamification of learning physics," *Gravity J. Ilm. Penelit. dan Pembelajaran Fis.*, vol. 6, no. 1, pp. 72–76, 2020, doi: 10.30870/gravity.v6i1.7108.
- [28] C. K. Meng, J. S. Mohd Nasir, T. M. Ming, and K. A. Choo, "A Gamified Classroom with Technical and Vocational Education and Training (TVET) Students using Quizziz," *Int. J. Educ. Islam. Stud. Soc. Sci. Res.*, vol. 4, no. 1, pp. 1–6, 2019, [Online]. Available: <http://ijeisr.net/Journal/Vol-4-No-1-Isu-02.pdf>.
- [29] G. Lückemeyer, "Virtual blended learning enriched by gamification and social aspects in programming education," *10th Int. Conf. Comput. Sci. Educ. ICCSE 2015*, no. Iccse, pp. 438–444, 2015, doi: 10.1109/ICCSE.2015.7250286.
- [30] C. Mese and O. O. Dursun, "Effectiveness of gamification elements in blended learning environments," *Turkish Online J. Distance Educ.*, vol. 20, no. 3, pp. 119–142, 2019, doi: 10.17718/tojde.601914.

# Campus Quality of Services Analysis of Mobile Wireless Communications Network Signal among Providers in Malaysia

Murizah Kassim<sup>1</sup>, Zulfadhli Hisam<sup>2</sup>, Mohd Nazri Ismail<sup>3</sup>

Institute for Big Data Analytics and Artificial Intelligence (IBDAAI)Universiti Teknologi MARA  
40450 Shah Alam, Selangor, Malaysia<sup>1</sup>

School of Electrical Engineering, College of Engineering, Universiti Teknologi MARA, 40450 Shah Alam, Selangor, Malaysia<sup>1,2</sup>  
Faculty of Science and Defence Technology, National Defence University of Malaysia, Kuala Lumpur, Malaysia<sup>3</sup>

**Abstract**—Wireless communication is very important in this generation where today's 5G internet connection is still unconfirmed and 4G communication is still needed. Network in Malaysia has been supported by many telecommunication companies and the Quality of Services is still poor supported especially in the campus area. This research presents a performance analysis of Quality of Services for 4G wireless Communication among Providers supported in a campus area in Malaysia. A 4G Nemo Outdoor wireless analyzer was used to collect the Reference Signal Received Power (RSRP) signal data based on the identified campus road maps. Digi and U-Mobile Network was identified and compared as two telecommunications providers in the testing. The identified road maps were analyzed along the routes while testing signals are collected while driving. It is identified that Digi supports better for the Mobile broadband network which shows an excellent of 1% and good connections of 29 % and 0% signal loss in the drive areas. RSRP signal for U-Mobile shows there is 8% signal loss and the connections provided only at the Mid-Cell for 43% and Cell Edge connections for 48%. This concludes that the 4G signal strength in the campus area having average signal strength, but some medium signal strength is also identified based on the road locations. This research is significant for QoS of supports mobile network in a campus area.

**Keywords**—Quality of services; 4G/LTE; mobile network; wireless communication; RSRP; campus network

## I. INTRODUCTION

Wireless communications have enabled the connection of billions of people to the Internet which they benefit from today's digital economy. A mobile phone is one of today's wireless communication devices where it allows people to use their devices and communicate everywhere in the world. It is identified that every sector of the economy now relies upon wireless technologies such as in banking, agriculture, transportation, healthcare, education, and many more [1]. Today's, the development of 4G to 6G services, mobile network, high-speed data, wireless sensor network and broadband services have become the most important sources of mobile communications operation services [2, 3]. Population changes using mobile phone data has been investigated and the social networking sites use and college students' academic performance on testing for an inverted U-shaped relationship using automated mobile app usage data

also has been analyzed[4]. 5G is a robust wireless communications networks but 5G is still in the stage of testing in Malaysia. Performance Analysis of Mobile Broadband Networks with 5G Trends and planning of antenna for future 5g energy harvesting in Malaysia has been done[5, 6]. Many applications and systems today like artificial intelligence and the Internet of Things need higher bandwidth to achieve QoS in communications either in wireless broadband or wireless sensor network [7, 8]. The advantage of using the wireless network is the costing is inexpensive compared to a wired network. A wireless network is using Radio Frequency (RF) for transmitting and receiving data by using the wave. The internet can be achieved on two platforms which are connections through coverage mobile data plan subscribed to the telecommunication by using the smartphone. Secondly are the broadband connections devices for the installation of mobile data for broadband providers needs updates checking for QoS. Some connections are identified as loss and slow. Thus, identify routes or areas of the supported broadband needs manually testing by using certain software to inspect all places without the reference map of routes. The other problems of data transmissions are noise.

This research described the performance analysis that has been done for QoS for 4G Wireless Communication among Providers which was tested in a campus area. A Nemo Outdoor tools and software has been used as the platform to analyze the transmission data 4G while driving test along the identified route maps in a campus network is identified for data collections. The Reference Signal Received Power (RSRP) performance of the 4G signal strength is presented to show the QoS for both providers in supporting coverage in the campus network.

## II. RELATED WORK

There are many types of performance analysis of the wireless network of 4G. Multiple-input and multiple-output (MIMO) is one technique for performance analysis for 4G Wireless. One research has presented that MIMO and orthogonal frequency division multiplexing (OFDM) were used for supported high data rate and high performance in different channel conditions [9]. Many methods have shown the use of different applications of voice and data connection

in the 4G wireless network. 4G analysis performance also used MADM algorithms which are integrated into three types of network situations. The network is WLAN, UMTS, and WiMAX [10]. Some references identified that the performance analysis for 4G wireless communication had three possible architecture types which are multimode devices, overlay networks, overlay networks, and common access protocol. The multimode device can access services in a different wireless network. The multimode device can improve call and expand effective coverage areas [11]. Conventional mobility management schemes tend to hit the core network with increased signaling load when the cell size is shrinking and the user mobility speed increases. A survey has been done for the idle mode mobility management and then proposes a new architecture, namely predictive mobility management (PrMM) to mitigate the identified challenges [12]. Malaysia has many telecommunications providers' services for wireless communication such as 4G/LTE. The famous providers in Malaysia are Celcom, Digi, Maxis, U-Mobile, and many more. These providers are much designated for the wireless network for the 4G services. 4G services are identified as the promise of a higher platform of a wireless network in the world although 5G has been implemented not all countries are ready and supported on the platform [13]. The fourth generation was upgraded from the three-generation 3G network. The different and the important of upgrading from the 3G to 4G network are the specifications on the coverage of speed and the costing to the consumers. The Table I shows the DIGI data features and Table II shows the U-Mobile data features on the download speed, upload speed, and latency between 3G and 4G of Digi and U-MOBILE network [14].

The keys of the 4G infrastructures are accessing information connected to a wide range of information and services, and receiving a large volume of information, data, pictures, voice, and video [15]. 4G is using the Orthogonal Frequency Division Multiplexing (OFDM) [16]. Besides, the advantages of using the 4G have advantages of supporting a higher speed that can reach up to 100Mbps. Using 4G with higher speed can do many things such as playing online games, watching high-definition video streaming, VOIP and can get interactive TV [17, 18]. The 4G there have 5 important ways and factors of making the 4G are Orthogonal Frequency Division Multiplexing (OFDM), Mobile WiMAX, Ultra Mobile Broadband (UMB), multiple-input multiple-output (MIMO), and Long-Term Evolution (LTE) [19]. It is mentioned that Mobile broadband (MBB) is one of the critical goals in fifth generation (5G) networks due to rising data demand. MBB provides very high-speed internet access with seamless connections. Existing MBB, including third-generation (3G) and fourth-generation (4G) networks, also requires monitoring to ensure good network performance [20]. Fig. 1 shows the Long-term Evolution or LTE are the norm for mobile device wireless broadband and Global Mobile

Communication System (GSM) information terminals. Using LTE can improve the information network's ability and speed focus applied in Malaysia[21]. Using LTE can enhance the capacity and speed of the data network. LTE characteristic are the bandwidths are can improve from 1.4 MHz until 20 MHz, LTE also supports the frequency division duplexing (FDD) and the time division duplexing (TDD)[18]. Additionally, the LTE can support the voice and data to the cell towers with older network technology such as CDMA2 2000 that are family of 3G mobile technology that can be sending voice, data, and signaling data besides mobile phone and cell sites [22].

TABLE I. DIGI DATA FEATURES

Feature	3G	4G
Download speed (Mb/s)	1.4	7.7
Upload speed (Mb/s)	0.3	3.4
Latency (ms)	657	45

TABLE II. U-MOBILE DATA FEATURES

Feature	3G	4G
Download speed (Mb/s)	1.7	13.3
Upload speed (Mb/s)	0.4	3.4
Latency (ms)	630	31

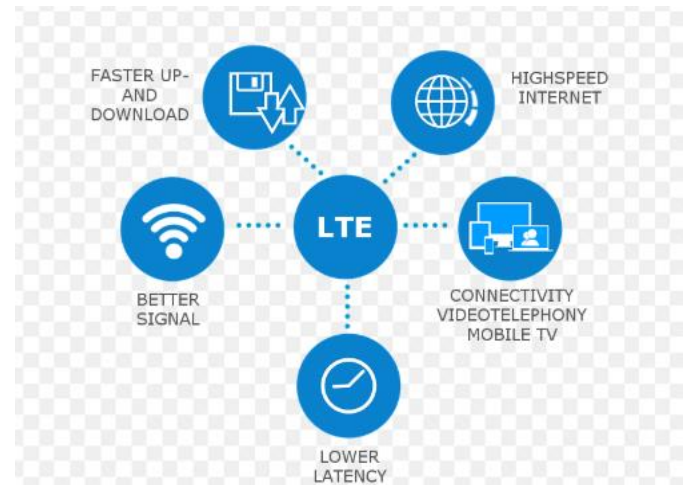


Fig. 1. Network Concept of Long-Term Evolution (LTE).

### III. PROPOSED METHOD

#### A. Research Flow

Fig. 2 shows the research flowchart of the process. There are a few steps that need to follow in collecting the data by using the Nemo Outdoor

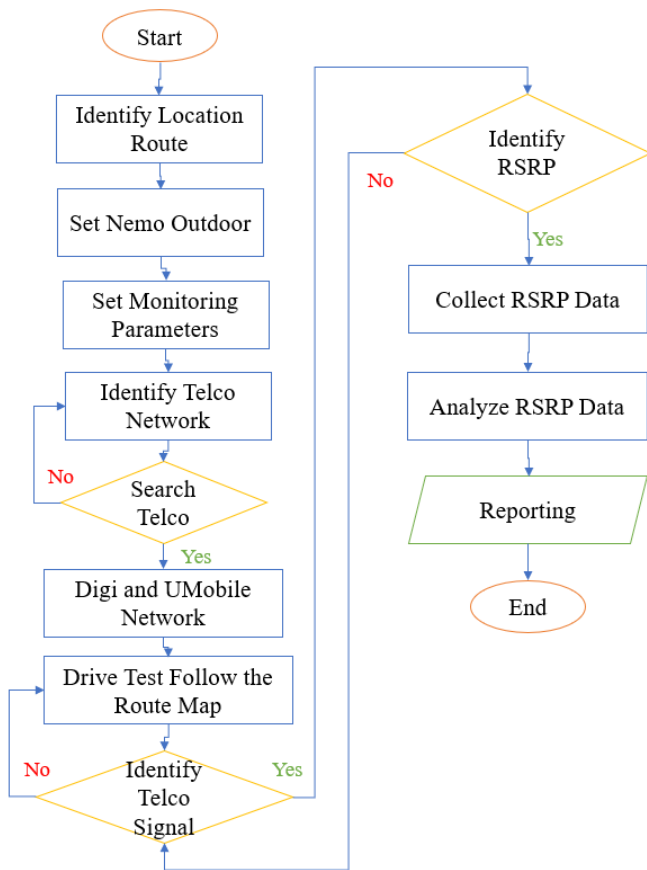


Fig. 2. Research Flowchart.

First, the task of checking or selecting the type of telecommunication network that can be identified around the campus area in Shah Alam is defined. Digi and U-Mobile were chosen based on the pre-trial where both signals gave the highest output in the communications signal. Next, the Nemo outdoor device and software is set for mapping signal is configured. After the configurations are justified and pretesting, the Nemo outdoor devices are to be ON while driving test followed by the route maps was tested. Data collections on the signal strength from Digi and U-Mobile have been set according to its frequencies on the Nemo Outdoor device. IF there are errors occurred then the driving test must be done again until the signal is collected successfully. During the drive test process, the data collection on the signal must be in real-time captured smoothly by time. Lastly, the result of the data collections is smoothly continued to be extracted from the Nemo software. Data is analyzed based on the defined wanted graph.

### B. Campus Route Map

The location of the route maps is defined to search for the RSRP Signal strength where setting is to be done on the Nemo outdoor. The Nemo Outdoor software then will open the maps to know the location of selected area for signal data collection. A Campus area has been selected which analysis on the data is important to present the reliability for students' connections in the campus who subscribed to the mobile data plan.

1) *Nemo monitoring tools*: The process of collecting the data is by drive testing which driver needs to ride the vehicle by following or marking the maps to get the data network signal and strength for RSRP. RSRP can be identified on the Long-Term Evolution (LTE) and 4G network. The admin for the data collector needs to insert a type of SIM card networks such as Digi or U-mobile into the smartphone of the Nemo Outdoor. This process is easier for data collection for each type of network, and it is read from the smartphone. Next, after the Handy Nemo software is activated from the admin or driver smartphone who drives the car according to routes of the campus map. The route of the campus maps also is mapping to the Nemo software. However, the route also can be created by the driver while driving the car. Table III shows the parameters for the used Nemo Outdoor analyzer.

TABLE III. DESCRIPTION OF NEMO OUTDOOR DEVICES

Type of Devices	Description
Laptop	<ul style="list-style-type: none"> <li>To open the software of the Nemo Outdoor</li> <li>To extract or transfer the data collect from the smartphone (Handy Nemo software)</li> </ul>
Smartphone	<ul style="list-style-type: none"> <li>Known as Handy Nemo devices in collecting the data network</li> </ul>
Dongle (Outdoor Measurement)	<ul style="list-style-type: none"> <li>The license to open the software of Nemo Outdoor</li> <li>To measure the measurement collection data</li> </ul>
Dongle (Outdoor Playback)	<ul style="list-style-type: none"> <li>The license to open the software of Nemo Outdoor</li> <li>To measure the data playback of the route</li> </ul>

Fig. 3 shows the handy Nemo Outdoor software with a specified setting and Fig. 4 shows the identified maps for data collections on the Digi network and U-mobile 4G network signals which have been linked to the laptop interface platform. Fig. 5 shows the Nemo Outdoor playback and measurement for the data signal collections.



Fig. 3. Handy Nemo Software.



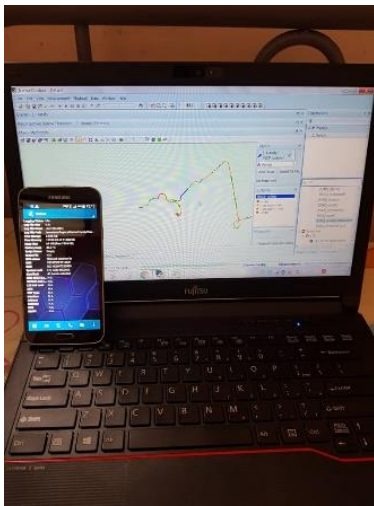


Fig. 4. Interface of Maps to PC or Mobile.



Fig. 5. Outdoor Playback & Measurement.

2) *Digi and U-Mobile network*: The Digi and U-Mobile network has been identified for analysis based on the best pretesting signals achieved compared to the other providers. The references for checking the quality and strength of the 4G network have been established. The performance analysis has measured the RSRP for both Digi and U-Mobile network. RSRP is the average Resource Elements (RE) power that carries cell-specific Reference Signals (RS) across the whole bandwidth. Thus, RSRP is measured only in the RS symbols. RSRP is the average of one RS resource element received. Fig. 6 and Fig. 7 shows the driving test followed the route of maps of in the campus area for both networks.

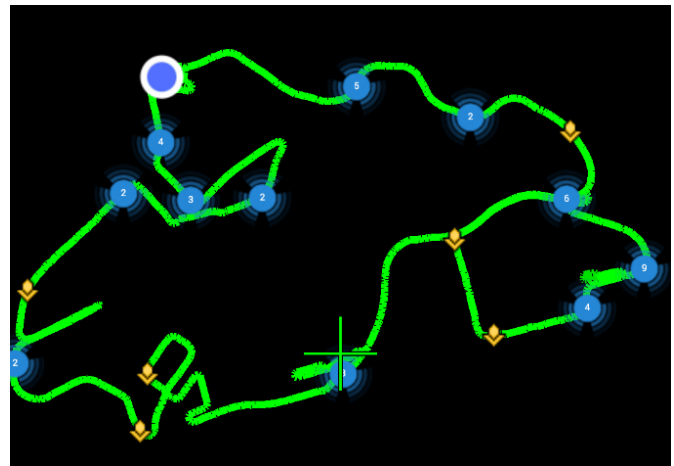


Fig. 6. Digi 4G Network Map.

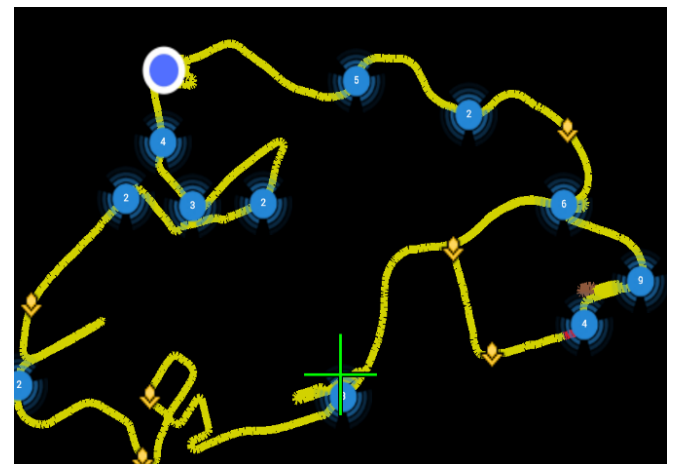


Fig. 7. U-Mobile 4G Network Map.

The Handy Nemo Outdoor has collected the data of each sector in the campus area in Shah Alam by following the routes map. The reading data on strength of the signal by the 4G network of Digi and U-Mobile are collected. The person in charge or driver needs to understand the condition of strength RSRP either it is Excellent, Good, Mid Cell, and Cell Edge while driving. Table IV shows the RF condition for the 4G signal strength. Additional quality can be measure by Reference Signal Received Quality (*RSRQ*), but it is not measured in this research.

TABLE IV. RADIO FREQUENCY (RF) CONDITION 4G SIGNAL STRENGTH

RF Condition	RSRP (dBm)	RSRQ (dB)
Excellent	$\geq -80$	$\geq -10$
Good	-80 to -90	-10 to -15
Mid Cell/ Medium	-90 to -100	-15 to -20
Cell Edge	$\leq -100$	$< -20$

3) *LTE signals strength*: RSRP is the average power received from a single Reference signal, and its typical range is around -44dbm (good) until -140dbm (bad) and the RSRQ is indicating the quality of the received signal, and its range is typical -19.5dB (bad) to -3dB (good).

- Identification of providers Signal

The identification of providers signal has been tested by using the Nemo Outdoor Software. This process needs to be done separately where the providers' signals will automatically be identified by the Nemo software. If the signals are undefined in the area the system will record as No Signal. Along the way while driving testing if the signal is changing the providers the system will record as Change Cells of network identification.

- RSRP Signal Strength

Nemo Outdoor Software is a multifunction for taking any data on wireless communication. The RSRP signal strength of the 4G network has been collected when the process identifies the provider's signal. The person in charge of the driving test needs to identify the signal strength of RSRP data collection. Some instructions must be followed to ensure the collection data signal strength of RSRP.

4) *Data Collection RSRP*: Two approaches can be done in collecting the data signals. First is the signal data can be request from the Research and Development team of each provider, but normally this approach is hard where data is confidential for outsiders. The second approach is where researchers must collect themselves the data based on the targetted area. Thus, Nemo Outdoor device and software is one of the most usable devices for data collections. Nemo outdoor provided data collections for RSRP strength of the 4G network. Few steps to be followed by the Nemo outdoor users for data collections in order the signal to be collected correctly without failed for data analyzing. The steps of setting up Nemo needs to be explored more where connections to the end devices like smartphone, laptop, and dongle of Nemo's license must be correct and properly running.

The research gaps have been identified such as the Performance, Speed, Frequency of 4G Network and RSRP Signal strength has been compared and identified in analyzing the performance of Campus Quality of Services Analysis of Mobile Wireless Communications Network Signal among Providers in Malaysia.

#### IV. RESULT AND ANALYSIS

##### A. RSRP Signal Strength

Fig. 8 and Fig. 9 show the data collection of Digi 4G Network and U-mobile 4G Network strength which is the RSRP signal for about an hour. The graph shows the condition strength of both networks was the same, but it has different

condition strength between each time in minutes. Result explained that most signals of LTE were greater than -100dbm, that means some of the area in the campus area have faced bad 4G/LTE signal of LTE.

##### B. Digi and U-Mobile RSRP Performance Analysis

The performance analysis for 4G/LTE signal strength is presented for both networks. The process of analyzing the collecting data which refers to the LTE signal strength is referred to Table IV. Result presents the 4G network signal strength based on the data referred to the reading of RSRP. The RF condition of the strongest signal was divided into four categories. First, Excellent that range is below and equal to -80dbm. Second, Good range from -80dbm until -90dbm. Third, Mid Cell range from -90dbm until -100dbm, and lastly cell edge was above -100dbm it shows that the signal strength was weak in those areas. Fig. 10 shows a five (5) hour data collection and Fig. 11 shows the best of RSRP signal strength that can be concluded that it has a few places that have a good 4G/LTE signal strength. The places that have the good signal strength of LTE were at the resident places which are the colleges, faculty, and cafeteria.

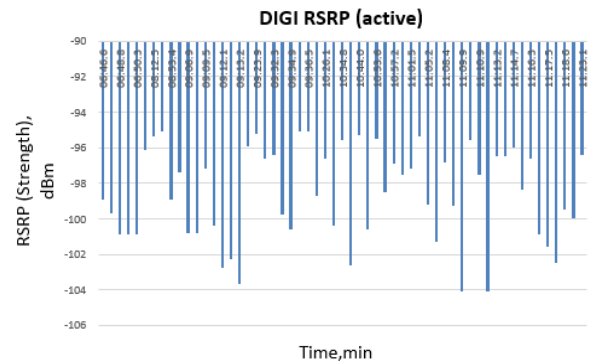


Fig. 8. RSRP Signal for Digi 4G Network.

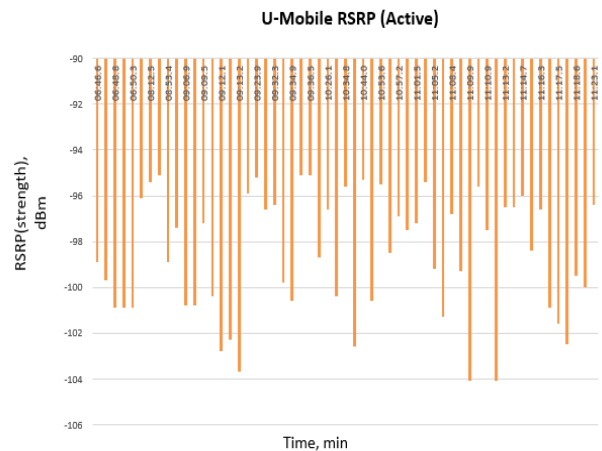


Fig. 9. RSRP Signal for U-Mobile 4G Network.



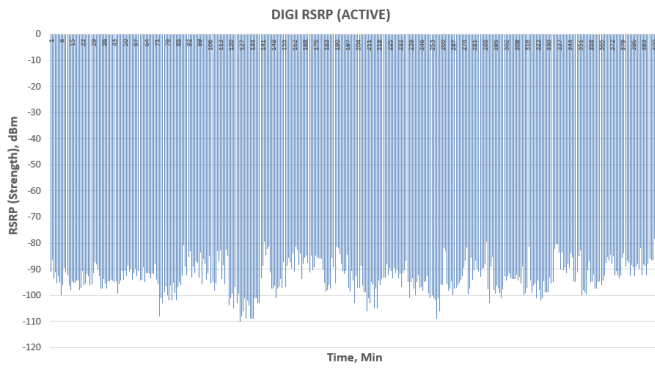


Fig. 10. Digi 4G/LTE Signal Strength Network.

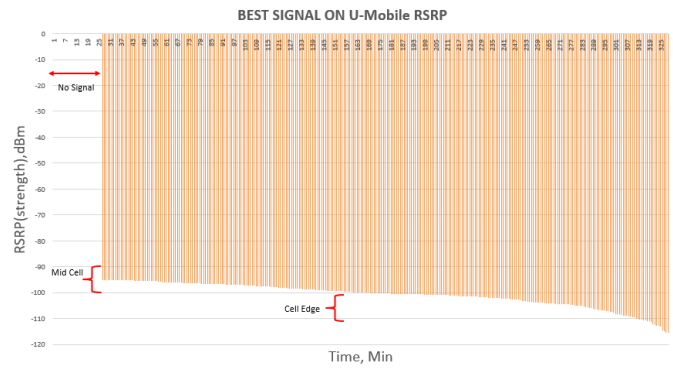


Fig. 13. Best U-Mobile 4G/LTE Signal Strength Network.

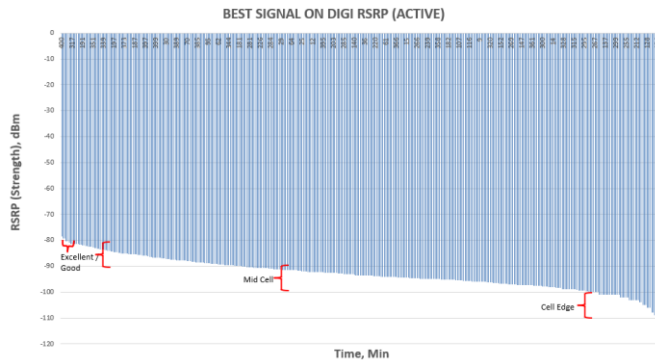


Fig. 11. Best Digi 4G/LTE Signal Strength Network.

Fig. 12 shows that some areas in the campus have a bad result of the signal strength of LTE or 4G network for U-mobile network. Some are having loss signal. Fig. 13 shows the best identified signal strength and most of the result from the signal strength in the campus area was only medium-range and cell edge signal strength. This is due to result shows the strength was mostly over and above from the -90dbm until -100dbm an above. Table V presents the comparison of Digi and U-Mobile RSRP identified form the research. It is identified that Digi supports better for the Mobile broadband network which shows an excellent of 1% and good connections of 29% and 0% signal loss in the drive areas. RSRP signal for U-Mobile shows there is 8% signal loss and the connections provided only at the Mid-Cell for 43% and Cell Edge connections for 48%.

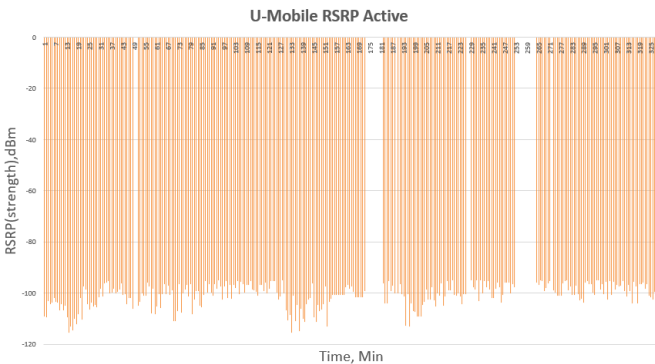


Fig. 12. U-Mobile 4G/LTE Signal Strength Network.

TABLE V. DIGI AND U-MOBILE COMPARISON PERCENTAGE OF RSRP

RSRP Strength	Digi	U-Mobile
Excellent	1%	0%
Good	29%	0%
Mid-Cell	59%	43%
Cell Edge	10%	48%
No Signal	0%	8%

DiGi comes top for Download Speed Experience which presents DiGi users experienced the fastest overall mobile download speeds in Malaysia. This result has supported the Mobile Network Experience Report that presents 17.6 Mbps on average of 8% (1.3 Mbps) faster than Maxis’s users, 37.5%-40.4% faster than U Mobile and Unifi users, and almost 84% faster than those on Celcom broadband mobile network [23].

## V. CONCLUSION

This research has successfully analyzed the Quality of Services for 4G Wireless Network Communications among Providers in a campus network. Two main providers were analyzed which is Digi and U-Mobile which are most identified signal in the campus network. The transmission data of the LTE or 4G signal strength is identified which shows how the transmission data was done using the Nemo Outdoor. The transmission of signal strength of 4G has shown a different strength in each different place in the campus between the providers. Research also identified that data collection to analyse the problem of the quality and the performance of the signal strength that needs to improve. The used of NEMO Outdoor is a good platform for company telecommunication to get the signal and quality strength of the 4G network which is easy to detect and analyze the place that has the low 4G network signal strength. Future research recommendation is to try to make a network that can detect more different types of data analysed.

## ACKNOWLEDGMENT

The author would like to thank the Institute for Big Data Analytics and Artificial Intelligence (IBDAAI), Universiti Teknologi MARA, 40450 Shah Alam, Selangor, Malaysia for the support fund in publishing the paper.

REFERENCES

- [1] R. Ma, K. H. Teo, S. Shinjo, K. Yamanaka, and P. M. Asbeck, "A GaN PA for 4G LTE-Advanced and 5G: Meeting the telecommunication needs of various vertical sectors including automobiles, robotics, health care, factory automation, agriculture, education, and more," *IEEE Microwave Magazine*, vol. 18, no. 7, pp. 77-85, 2017.
- [2] Y. Ding, "Retention Strategy for Existing Users of Mobile Communications," in *International Conference on Cognitive based Information Processing and Applications (CIPA 2021)*, 2022: Springer, pp. 310-317.
- [3] G. Al-Mamari, F. Bouabdallah, and A. Cherif, "A Survey of Sink Mobility Models to Avoid the Energy-Hole Problem in Wireless Sensor Networks," *International Journal of Advanced Computer Science and Applications*, Article vol. 13, no. 5, pp. 981-993, 2022, doi: 10.14569/IJACSA.2022.01305110.
- [4] W. Tafesse, "Social networking sites use and college students' academic performance: testing for an inverted U-shaped relationship using automated mobile app usage data," *International Journal of Educational Technology in Higher Education*, Article vol. 19, no. 1, 2022, Art no. 16, doi: 10.1186/s41239-022-00322-0.
- [5] I. Shayea et al., "Performance Analysis of Mobile Broadband Networks with 5G Trends and Beyond: Urban Areas Scope in Malaysia," *IEEE Access*, Article vol. 9, pp. 90767-90794, 2021, Art no. 9446158, doi: 10.1109/ACCESS.2021.3085782.
- [6] A. K. M. Zakir Hossain, N. B. Hassim, S. M. Kayser Azam, M. S. Islam, and M. K. Hasan, "A planar antenna on flexible substrate for future 5g energy harvesting in Malaysia," *International Journal of Advanced Computer Science and Applications*, Article vol. 11, no. 10, pp. 151-155, 2020, doi: 10.14569/IJACSA.2020.0111020.
- [7] A. A. Ghafar, M. Kassim, N. Ya'acob, R. Mohamad, and R. A. Rahman, "Qos of wi-fi performance based on signal strength and channel for indoor campus network," *Bulletin of Electrical Engineering and Informatics*, Article vol. 9, no. 5, pp. 2097-2108, 2020, doi: 10.11591/eei.v9i5.2251.
- [8] S. Pandey and G. R. Kadambi, "Modeling Wireless Mesh Networks for Load Management," *International Journal of Advanced Computer Science and Applications*, Article vol. 13, no. 5, pp. 241-251, 2022, doi: 10.14569/IJACSA.2022.0130530.
- [9] A. Idris, A. N. Farhana, H. Adiba, and M. Kassim, "BER and PAPR analysis of MIMO OFDMA and SCFDMA system using different diversity techniques," in *Proceedings - 7th IEEE International Conference on Control System, Computing and Engineering, ICCSCE 2017*, 2018, vol. 2017-November, pp. 293-298, doi: 10.1109/ICCSCE.2017.8284422.
- [10] R. Ab Rahman, M. Kassim, Y. Cik Ku Haroswati Che Ku, and M. Ismail, "Performance analysis of routing protocol in WiMAX network," in *Proceedings - 2011 IEEE International Conference on System Engineering and Technology, ICSET 2011*, 2011, pp. 153-157, doi: 10.1109/ICSEngT.2011.5993440.
- [11] T. A. Benmusa, A. J. Belgassem, and M. A. Ibrahim, "Planning and dimensioning a high speed 4G WiMAX network in Tripoli area," in *14th international conference on Sciences and Techniques of Automatic control & computer engineering-STA'2013*, 2013: IEEE, pp. 318-324.
- [12] S. A. Hoseinitabatabei, A. Mohamed, M. Hassanpour, and R. Tafazolli, "The Power of Mobility Prediction in Reducing Idle-State Signaling in Cellular Systems: A Revisit to 4G Mobility Management," *IEEE Transactions on Wireless Communications*, vol. 19, no. 5, pp. 3346-3360, 2020.
- [13] T. Mumtaz, S. Muhammad, M. I. Aslam, and N. Mohammad, "Dual Connectivity-Based Mobility Management and Data Split Mechanism in 4G/5G Cellular Networks," *IEEE Access*, vol. 8, pp. 86495-86509, 2020.
- [14] M. A. Hajar, D. N. Ibrahim, M. R. Darun, and M. A. Al-Sharafi, "Value innovation activities in the wireless telecommunications service sector: A case study of the Malaysian market," *Journal of Global Business Insights*, vol. 5, no. 1, pp. 57-72, 2020.
- [15] W. Tashan, I. Shayea, S. Aldirmaz-Colak, M. Ergen, M. H. Azmi, and A. Alhamadi, "Mobility Robustness Optimization in Future Mobile Heterogeneous Networks: A Survey," *IEEE Access*, Article vol. 10, pp. 45522-45541, 2022, doi: 10.1109/ACCESS.2022.3168717.
- [16] A. Idris, N. A. M. Deros, I. Taib, M. Kassim, M. D. Rozaini, and D. M. Ali, "PAPR reduction using huffman and arithmetic coding techniques in F-OFDM system," *Bulletin of Electrical Engineering and Informatics*, Article vol. 7, no. 2, pp. 257-263, 2018, doi: 10.11591/eei.v7i2.1169.
- [17] M. Njikam, S. H. Nanna, S. Shahrin, and M. F. I. Othman, "High speed internet development in Africa using 4G-LTE technology-a review," *Bulletin of Electrical Engineering and Informatics*, Article vol. 8, no. 2, pp. 577-585, 2019, doi: 10.11591/eei.v8i2.684.
- [18] M. Kassim, R. A. Rahman, M. A. A. Aziz, A. Idris, and M. I. Yusof, "Performance analysis of VoIP over 3G and 4G LTE network," in *2017 International Conference on Electrical, Electronics and System Engineering, ICEESE 2017*, 2018, vol. 2018-January, pp. 37-41, doi: 10.1109/ICEESE.2017.8298391. [
- [19] I. Shayea, M. H. Azmi, T. A. Rahman, M. Ergen, C. T. Han, and A. Arsad, "Spectrum gap analysis with practical solutions for future mobile data traffic growth in Malaysia," *IEEE Access*, vol. 7, pp. 24910-24933, 2019.
- [20] A. A. El-Saleh et al., "Measuring and assessing performance of mobile broadband networks and future 5G trends," *Sustainability*, vol. 14, no. 2, p. 829, 2022.
- [21] Y. J. I. A. Yghoubi, W. L. Pang, S. K. Wong, and K. Y. Chan, "Performance evaluation of video streaming on LTE with coexistence of Wi-Fi signal," *Bulletin of Electrical Engineering and Informatics*, Article vol. 8, no. 3, pp. 890-897, 2019, doi: 10.11591/eei.v8i3.1580.
- [22] S. Mathur et al., "Demo abstract: CDMA-based IoT services with shared band operation of LTE in 5G," in *2017 IEEE Conference on Computer Communications Workshops, INFOCOM WKSHPs 2017*, 2017, pp. 958-959, doi: 10.1109/INFCOMW.2017.8116509.
- [23] Hardik Khatri, "Mobile Network Experience Report," in *National Analysis*, ed: Opensignal Limited, 2022, pp. 1-5.

# Exploring Alumni Data using Data Visualization Techniques

Nurhanani Izzati Ismail, Nur Atiqah Sia Abdullah\*, Nasiroh Omar

Faculty of Computer and Mathematical Sciences  
Universiti Teknologi MARA  
Shah Alam, Selangor, Malaysia

**Abstract**—Alumni data are mostly managed through paper-based and word file. With lots of alumni graduating each year, these massive data become difficult to handle. It is hard to look for past alumni data to know their current situation. Since the data is kept conventionally, there is no communication between the alumni and the faculty. Therefore, we proposed a solution that includes alumni information regarding their status in life where alumni themselves and individuals in the faculty can see. This study aims to visualize alumni data from a faculty in a public university through an exploratory dashboard using the identified data visualization techniques. This study adopts the dashboard development process consists of three major phases, which are conception, visualization, and finalization phase. The primary audience is identified and the theme for the dashboard is decided in the conceptual phase. The primary and support views are then designed together with the layout during the visualization phase. At the end of the study, the exploratory dashboard for alumni data using multidimensional and hierarchical data visualization is finalized with the interactive elements. The results are interpreted through descriptive and diagnostic analysis. The dashboard is then evaluated through convenience sampling technique to verify the representation of the dashboard. Majority respondents agreed on the simplicity of exploratory dashboard and the amount of data is also sufficient with the selection of the visualization types. The dashboard is beneficial to the university's administrator, alumni, and public.

**Keywords**—Alumni; descriptive analysis; diagnostic analysis; data visualization; exploratory dashboard

## I. INTRODUCTION

Alumni is classified as a significant secondary source of revenue [1]. Alumni are one of the important resources for the university. They are the individuals who symbolize the university within the world. Numerous alumni networks were at first began from regional groups of alumni united for university fundraising activities. These associations gradually obtained an added significance within the evolution of the university due to their tremendous advocacy capability that assists the university and prepares students in their future profession.

By connecting with alumni, a university can keep on profiting from their abilities and knowledge. Alumni management is straightforwardly related to the amplification impact of alumni resources in this technology era. The issue is that the conventional alumni data management system is hard to adjust to the large alumni groups and the massive amount of data [2]. The current method used by the university in

managing alumni data is through the traditional way where the alumni management staff is maintaining paper-based documents to store the information, for example, alumni and college details.

Thus, it is hard to sustain the past information and important data through a paper-based method. Staffs require more time to create required alumni reports. It is dreary to administer past information, which requires much room to store every one of the past year's records and documents paperwork. For example, Universiti Teknologi MARA is among the public university in Malaysia and has branches in every state in the country. Thus, they have a lot of alumni every year and without data visualization, it will be hard to trace the alumni current situation or where they are going after they finish their study.

There are several universities in Malaysia that embodied data visualization in their systems for alumni, but the visualization is limited to certain criteria. This paper proposes an alumni data visualization through an exploratory dashboard using identified data visualization techniques. This study uses the data from alumni of the Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Malaysia. The data are the alumni general academic program data and employment data, without containing any personal information. These alumni data are visualized using multidimensional and hierarchical data visualization.

The paper consists of five sections. It starts with Section I Introduction; continues with Section II Related Works. It follows by Section III Methodology to report on the methods used in this study. Then, Section IV includes the results and discussion, and the last section concludes the study.

## II. RELATED WORKS

### A. Data Visualization

Data visualization is the act of presenting data or information into a pictorial or graphical context, such as a chart, a map, or other visual formats, to make information uncomplicated for the human brain to comprehend and understand [3].

Data visualization is a sector that corporates data from numerous fields, for example, psychology which examines the data insight and the effect of certain components on data insight, next is the computer science and statistics field which expanded ideas in the new area like artificial intelligence and

data mining methods [4]. The visual and multimedia design field is also crucial to assembling infographic dashboards which involve a few elements including data, scales, lines, bars, and colored shaped sizes.

**B. Types of Data Visualization**

Some of most popular data visualization types are temporal, hierarchical, network, multidimensional, and geospatial data visualization. These data visualization types have different criteria and strengths in representing the alumni data.

1) *Temporal Data Visualization*: Data that have progressive values significant and recorded. Temporal visualizations usually include lines that either remains solitary or overlap with one another, with a beginning and end time [5]. Examples of temporal data visualization are scattered plots, line graphs, and polar area diagrams. This type of data visualization can represent the changes of data in timeline form. The parameters used in the alumni visualization are not suitable for temporal data visualization because there are no data on time-related. Thus, this study does not use the temporal data visualization.

2) *Hierarchical Data Visualization*: Information that can be organized as a tree [6]. The connection between the parent nodes and the child nodes forms a tree network. It shows how information was positioned and arranged together in a system [7]. Examples of hierarchical data visualization are treemap charts, tree diagrams, and sunburst diagrams. This type of data visualization can be used in data consisting of main categories and sub-categories. This study use a treemap chart to visualize alumni job category based on their domain.

3) *Network Data Visualization*: A set of nodes with links interfacing with the nodes [8]. Nodes stand for data points, and links represent the associations between them [9].

Examples of network data visualization are matrix charts, node-link diagrams, and word clouds. This data visualization is widely used on social media platforms such as Twitter and Instagram. Since the alumni data in this study has no relationship between parameters, network data visualization is not suitable to be used as the alumni data do not involve any associations among the dataset.

4) *Multidimensional Data Visualization*: It analyzes various data dimensions or qualities [10]. Multidimensional data visualization includes simply looking at distributions and possible connections, patterns, and relationships among these qualities. Examples of multidimensional data visualization are bar charts, Venn diagrams, and pie charts. Among all the five types of data visualization, multidimensional is commonly used for the dataset because the alumni data consists of many parameters that are suitable for multidimensional as it allows representation of multiple categories in the dataset.

5) *Geospatial Data Visualization*: It is the earliest type of data visualization [11]. Geospatial data visualization covers factors on a map using latitude and longitude to encourage understanding. Examples of geospatial data visualization are choropleth maps, cartograms, and heat maps. This data visualization is used for data that has locations and involve the use of the map. In alumni data visualization, choropleth is suitable to visualize alumni location because it uses values on a specific region on the map. However, we exclude this location information due to the privacy issue.

From Table I, two types of data visualization are more suitable to represent the relationship between alumni datasets in this study. Multidimensional data visualization and hierarchical data visualization can visualize multiple data variables such as alumni general data and alumni employment data, and it is easier to interpret the data and extract the information from the data.

TABLE I. COMPARISON OF DATA VISUALIZATION

Characteristic	Types of Data Visualization				
	<i>Temporal</i>	<i>Hierarchical</i>	<i>Network</i>	<i>Multidimensional</i>	<i>Geospatial</i>
<b>Type of data</b>	Time series, the event sequences	Data which have main category and sub-category	Relationship between nodes and links	Multiple data parameters and categories	Data involving geographical location
<b>Usage</b>	Use in climate data, and historical presentation	Use in politics, mathematics, or organizations	Use in social media platform	Use in medicine and social sciences	Use in Google maps, and weather maps
<b>Technique</b>	Scatter plot, Line graph, Polar Area, Time Series	Tree Diagram, Sunburst Diagram, Tree Map Chart, Circle Packing	Matrix Chart, Node-Link Diagram, WordCloud, Sankey Diagram	Stacked Bar Chart, Pie Chart, Venn Diagram, Histogram	Choropleth Map, Density Map, Cartogram, Heat Map
<b>Justification for the usage of data visualization in the proposed alumni data visualization</b>	The parameters use in the alumni visualization is not suitable for temporal data visualization because there are no time-related data.	Can be used in data consists of main category and sub-category. This study uses treemap chart to visualize alumni job category based on their domain	<i>The alumni data used in this study are simple data, network data visualization is not suitable to be used as the alumni data do not involve any associations among the datasets.</i>	Multidimensional is commonly used because the alumni data consists of parameters that allows representation of multiple categories in the dataset.	The alumni data used in this study is not suitable for geospatial data visualization because there are no data on alumni state of origin.

C. Alumni Data Visualization

This section focuses on analyzing five existing alumni data visualization models from foreign and Malaysia local universities.

1) *Graduate Alumni Salaries – University of Colorado Boulder*: It develops a data visualization for its alumni based on their salaries as shown in Fig. 1. This university used multidimensional, geospatial and hierarchical data visualizations [12]. The multidimensional data visualization is horizontal bar charts to visualize the alumni salaries, the job category and the employer category. The geospatial data visualization is a choropleth map to show the employer location by state. A treemap, which is a hierarchical data visualization, shows the top industries category where it displays categories similar to each other and any correlations between them. This data visualization has a filter feature to enable the user to search data based on specific keywords. This data visualization does not use temporal and network data visualization as it does not have any parameters that are suitable for that type of data visualization. This data visualization uses three colors, which are chocolate, grey and black. The use of three colors give a simpler and neat look. The visualization does not look crowded and messy.

2) *Mapping Institutes of Fine Arts (IFA), New York University Alumni*: It develops a data visualization for their alumni, which shows the current employment status of their active alumni as shown in Fig. 2 and Fig. 3 [13]. This institute uses network and multidimensional data visualizations. The network data visualization used is a network map that shows the current employment location of alumni where it shows how the different location is interconnected from one node to other nodes through link lines. The multidimensional data visualization used is a vertical bar chart to visualize the type of employment of the alumni. In this data visualization, several colors are used to represent the employment categories and display them in the bar graph. This data visualization does not use geospatial, temporal or hierarchical data visualization because the dataset involved is not related to this type of data visualization. This data visualization also uses a filter feature to enable the user to search alumni data based on institutes and people.

3) *Alumni of Universiti Malaysia Sarawak*: The university developed a data visualization for their alumni, which shows the number of alumni graduates from each faculty from 2006 to 2015 [14]. This data visualization used multidimensional data visualization and display data insight for the website visitor as shown in Fig. 4. The university used a stacked bar graph to show the trend of alumni graduates based on their level of study. This data visualization does not use an interactive dashboard. Thus, the visitor can only view the website for the alumni data and they cannot interact with the data. This data visualization only uses multidimensional. Thus, the presentation of data is not complete as it only visualizes the number of alumni who graduated. The alumni visualization developed in this research study can enhance this

issue by implementing the interactive dashboard for the alumni and incorporating other types of visualization to represent different types of alumni data.

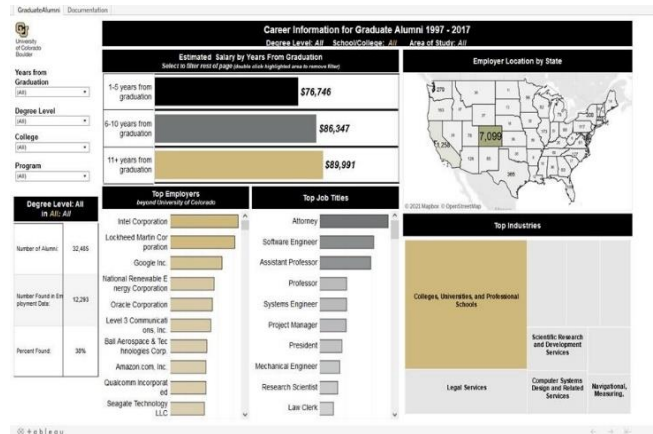


Fig. 1. Data Visualization for Graduate Alumni Salaries.



Fig. 2. Network Data Visualization for Institutes of Fine Arts.



Fig. 3. Multidimensional Data Visualization for Institutes of Fine Arts.



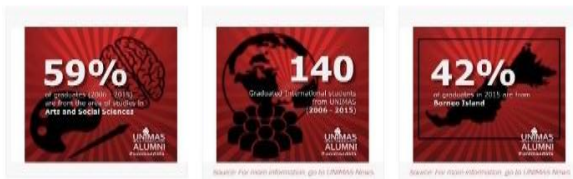
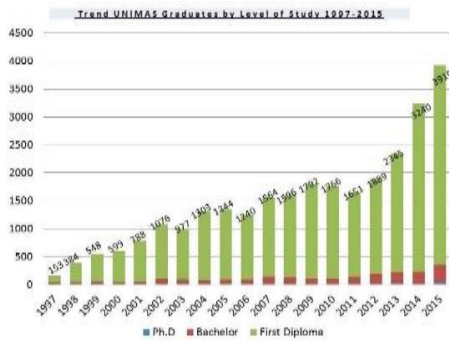


Fig. 4. Data Visualization for Universiti Malaysia Sarawak Alumni.

4) Alumni of Universiti Islam Antarabangsa Malaysia:

The univeristy has developed a data visualization for its alumni, which shows the total numbers of local and international alumni [15]. This interactive data visualization used geospatial data visualization to display the number of alumni based on the country of origin as shown in Fig. 5. It used a choropleth map to visualize the geographical area where it will display the number of alumni based on the state. The university does not use an interactive dashboard for their data visualization as they only implement visualization for alumni state. The visitor can hover over the map and the number of alumni is displayed for the country chosen. Since this data visualization only uses geospatial, a lot of alumni data are not shown and the visualization looks simple. The visualization can be enhanced by including more types of data visualization so that more alumni data can be visualized to user.

two types of data visualization in order to visualize the alumni general data and alumni occupation data.

From Table II, the existing data visualization developed by the University of Colorado Boulder shows the best data visualization method and suitable to be adopted in this study. This is because the data visualization used by the university is multidimensional data visualization, hierarchical data visualization and geospatial data visualization. In this study, we adopt this technique to visualize alumni general data and alumni employment data.



Fig. 5. Data Visualization for Universiti Islam Antarabangsa Malaysia Alumni.

5) Alumni Data of Heritage College of Osteopathic Medicine, Ohio University:

The university developed a data visualization for their alumni, displaying the percentage of alumni practising area and specialization [16]. This interactive data visualization used multidimensional and hierarchical data visualization to visualize the alumni data as shown in Fig. 6 and Fig. 7. A multidimensional data visualization used in this system is the pie charts to show the percentage of alumni practicing and training areas in the primary and non-primary sectors. While the hierarchical data visualization used is a treemap chart to show the number of alumni based on their specialization. In the pie chart, several colors are used to represent the keyword for the alumni practicing sectors. The pie chart has also been divided into four sections with four different colors to enable the user to see the data explicitly and differentiate it. The treemap chart also been divided into four different categories for the alumni specialization. This research study will include these

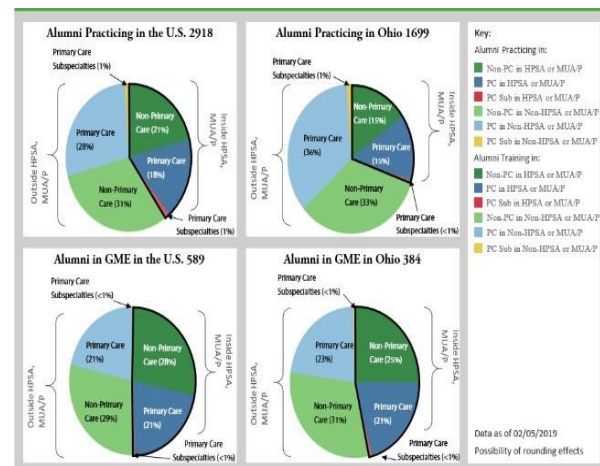


Fig. 6. Multidimensional Data Visualization for Ohio University Alumni.

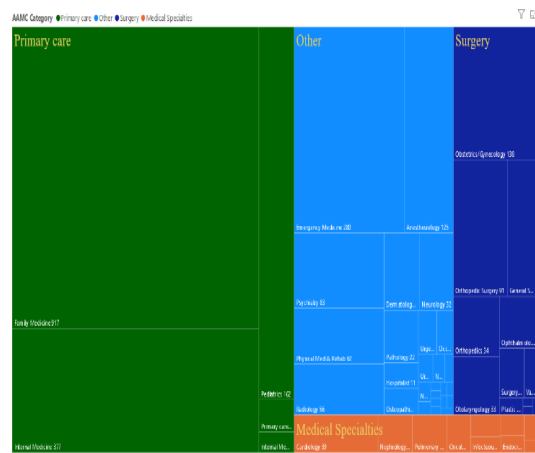


Fig. 7. Hierarchical Data Visualization for Ohio University Alumni.



TABLE II. COMPARISON OF EXISTING MODEL OF ALUMNI DATA VISUALIZATION

Characteristic	Existing Model of Alumni Data Visualization				
	<i>University of Colorado Boulder</i>	<i>Institutes of Fine Arts (IFA), New York University</i>	<i>Universiti Malaysia Sarawak (UNIMAS)</i>	<i>Universiti Islam Antarabangsa Malaysia (UIAM)</i>	<i>Heritage College of Osteopathic Medicine, Ohio University</i>
Category of data visualization	Multidimensional, Hierarchical and Geospatial data visualization	Network and Multidimensional data visualization	Multidimensional data visualization	Geospatial data visualization	Multidimensional and Hierarchical data visualization
Technique	Bar graph, treemap chart and Choropleth map	Network map and bar graph	Stacked bar graph	Choropleth map	Pie chart and Treemap
Strength	Search and filter function to enable user to search for specific data easily, Use of simple colors that make the dashboard look neat and not messy	Several colors are used to represent the employment categories and display in the bar graph and use of search and filter function	University used stacked bar graph to show the trend of alumni graduates based on their level of study	Used choropleth map to visualize the alumni state. Visitor can hover over the map and the number of alumni is displayed for the country chosen	The pie chart divided into four sections with different colors to enable user to see the data explicitly. The tree map chart divided into four different colors to indicate the different alumni specialization
Drawback	No use of network and temporal data visualization because no data related to time and association.	Do not use geospatial, temporal and hierarchical data visualization because the dataset involved is not related to this type of data visualization	Do not use an interactive dashboard. Thus, the visitor can only view the website for the alumni data, and they cannot interact with the data. This data visualization only uses multidimensional. Thus, the presentation of data is not completed	Do not use an interactive dashboard for their data visualization. Use geospatial data visualization to visualize alumni state, many alumni data are not shown, and the visualization looks simple	Do not use geospatial, temporal and network data visualization because the dataset involved is not related to this type of data visualization
Comparison to research topic	The use of the same type of data visualization to visualize the alumni data but with more parameters and categories	The use of the multidimensional data visualization for several data types in the alumni dataset	Alumni visualization developed in this research study can enhanced this issue by implementing the interactive dashboard for the alumni and incorporate other types of visualization to represent different types of alumni data	Inherit the geospatial data visualization from the university but including more type of data visualization so that more alumni data can be visualized to user	Include the same types of data visualization to visualize the alumni general data and alumni occupation data.

This study uses multidimensional data visualization the most because the dataset used has many parameters. This study uses the same technique to visualize alumni employment data based on their salary range. The study also uses hierarchical data visualization to show alumni job domain. The data visualization also enables the search and filtering function.

### III. METHODOLOGY

This study aims to develop an interactive dashboard with the visualization of alumni data for the alumni. It adopts the dashboard development process throughout the development phase [17]. The dashboard development process consists of three main phases, which are the conception phase, visualization phase and finalization. In this study, there are three types of alumni data that will be used to develop the dashboard for alumni visualization. The alumni data include alumni general data (gender, age range, and program name), and alumni employment data (job domain and salary range).

#### A. Conception Phase

There are three main activities during this phase. The first activity is to identify the primary audience for dashboard. The second activity is the main questions. The third activity is the theme.

- Identify primary audience: In this study, the primary end-user audience for the interactive dashboard is the alumni management staff of the university. The alumni of the university, parents, and public can also view the exploratory dashboard.
- Main questions: The questions focus on the purpose of the dashboard development and what the dashboard can do to answer the user's questions. Nine questions serve as the main ideas for the exploratory dashboard. The list of questions ensures that an exploratory dashboard can be developed and delivered based on these ideas. The list of questions is created based on several websites.
- Theme: this study aims to develop the exploratory dashboard, which uses the types of data visualization to answer the main questions. Alumni general data will be visualized by using hierarchical and multidimensional data visualization. Alumni employment data will be visualized using multidimensional data visualizations.

#### B. Visualization Phase

Three main elements need to be considered during this phase. The first element is the primary views. The second

element is the support views. The third element is the layout. During this phase, a prototype for the alumni data visualization is designed.

- Primary views: The primary views are views that visually address the main questions gathered for the specific users and align with the theme that has been selected.
- Support views: The support views are the contributory or helpful views that support, refine or add context to the primary views. For instance, the search or filtering function where the dashboard enables users to engage with the dashboard by enabling searching for alumni by state or filtering the alumni data by years of graduation or by salary range.
- Layout: It is a placement and alignment of views that focus users' attraction on the primary views and supporting views placed around them and visual indicators highlighting how the support views interact with primary views.

### C. Finalization Phase

In the finalization phase, there are two main activities taking place. The first main activity is the interactive elements. The second activity is perfect and feeling. At the end of this phase, evaluation is carried out through convenience sampling to evaluate the exploratory dashboard based on certain factors.

- Interactive elements: The interactions between the primary views and the supporting views are set up in a logical, progressively detailed sequence during this activity.
- Perfect and feel: The alignments, the fine-tune color, fonts, and fonts consistency will be finalized and ensure adherence to visual standards.

## IV. RESULTS AND DISCUSSION

This study developed an exploratory dashboard of alumni data visualization for Universiti Teknologi MARA. The alumni data visualization is illustrated in several diagrams of the preliminary design using Power BI.

### A. Pre-Processing in Power BI

The secondary data has approximately 714 alumni from three cohorts including graduate from March 2020, November 2020, and June 2021. The cleaning process filtered private and personal information. After that, we examine the dataset to discover any issues on the data that can lead to incorrect analysis and avoid from showing the irrelevant result.

In this study, there are two major tables for the alumni dashboard visualization. These two tables consist of alumni data for cohorts 2020 and 2021. The chosen columns are year of graduation, age, gender, program name, job domain, and salary range. From the columns, multidimensional data visualization is used to visualize alumni age range, gender, program name, and salary range. Meanwhile, hierarchical data visualization is used to visualize alumni job domains. This alumni data visualization embedded filtering function by using slicers.

The data columns in the slicers are year of graduation, gender, salary range and the job domain. Before creating the exploratory dashboard, a theme for the alumni data visualization has been decided and applied to ensure that all report items or panels are consistent. In this study, the alumni exploratory dashboard uses a blue and grey colour as an overall theme for the dashboard. The exploratory dashboard has two retrospective pages. The exploratory dashboard is specific to overview alumni data and their employment details.

### B. Descriptive Analysis

Fig. 8 shows the visualization of the alumni details based on the total number of alumni, the gender, the age range, the program name, and the job domain for the alumni.

The top of the dashboard displays the total number of alumni. About 500 of them are female alumni while 214 alumni are male. The data for the gender of alumni are represented using a pie chart that communicates a part-to-whole connection within the data. The age range is represented in the format of a horizontal bar graph. Alumni's ages ranged from 24 to 30 years old with 70.03% identified as female while another 29.97% identified as male alumni.

There are 10 programmes offered in this dashboard and represented using vertical bar graph. The program with highest number of alumni is Bachelor of Actuarial Science with 164 alumni and majority are female. The second highest number of alumni is Bachelor of Information Technology with 106 alumni.

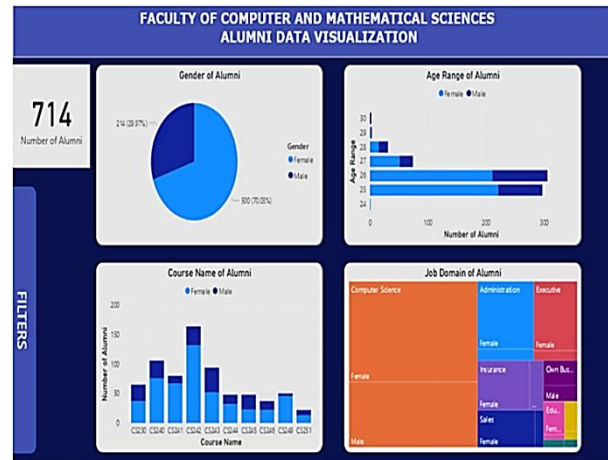


Fig. 8. The Visualization of Alumni General Data.

In terms of job domain, the data is represented in the form of a tree map chart. Most of the alumni were successfully employed in their respective field of study which is computer science with 56.3% which is equivalent to 402 alumni. Most of the alumni who are working in the computer science domain work as system analysts, programmers, and web developers. However, some of the alumni are working in other sectors that are not related to their field of study.

Fig. 9 shows the visualization of alumni salary range in six categories based on the job domain. The data are visualized through multidimensional data visualization by using horizontal bar graph. For the job domain in computer science,

most of the alumni have a salary range from RM2501 to RM3000. Majority of the alumni are working as system analysts, web developers and programmers. The job domain with the highest number of alumni is in the administration field that a salary ranging from RM1001 to RM1500. It is a challenge to the university to ensure that the graduates can be employed in the related field to their studies.

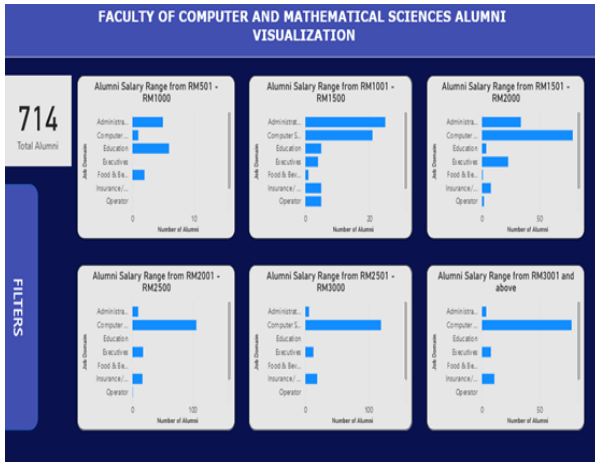


Fig. 9. The Salary Range of Alumni based on their Program.

### C. Diagnostic Analysis

In this study, there are several anomalies and abnormal occurrences from the alumni data visualization. There are several alumni work in the job domain that are irrelevant to their field of study in the university. For example, aside from the computer science field, most of the alumni are working in the administration line of work as admin assistants, executive, salespeople, operator, and teachers.

Most occupations in computer science field require specific skills such as knowledge in multiple programming languages and proficiency in coding a software program. Some alumni may feel not confident to pursue their career in the computer science field. Thus, they change their interest to work in the field that they are comfortable with. The observation result is aligned with the finding from the study done by Salas-Velasco [18] on education-job mismatch in the labour market for graduates of universities.

Secondly, there is an anomaly in the alumni salary's visualization. Unfortunately, some alumni are working in computer science with a low salary of RM1500 and below. According to recruiting businesses and analysts from the article reported by Institute of Strategic and International Study (ISIS) Malaysia on their website, fresh graduates in computer science have low beginning salaries because they lack digital skills in an increasingly competitive market and an uncertain economy [19].

Some of the software businesses that hire these fresh graduates are usually start-up companies that unable to offer a high beginning salary for the fresh graduates. With the high competition for employment in gigantic computer science companies in Malaysia, some alumni have no choice but to accept the offer from the small company with the lowest salary paid.

### D. Evaluation

After developing the exploratory dashboard for the alumni data visualization, an evaluation is done by conducting a convenience sampling method using survey form.

The alumni data visualization is evaluated based on the simplicity, choice of visualization, layout, filtering features, and view of the exploratory dashboard in mobile phone. Majority respondents agreed on the simplicity of exploratory dashboard. They can interpret easily. The amount of data is also sufficient as there are no data overload that led to messy representation of the alumni data visualization. Most respondents also agreed on the selection of the visualization types. However, they suggested the change of color to a brighter tone. Most of the respondents agreed and were satisfied with the arrangement of the layout for the visualization.

The respondents suggested to enlarge the font size to ease the view the information from dashboard. Most respondents agreed that the filtering function is useful as it enables them to navigate throughout the data visualization and locate the specific data and information. The filtering function encourages the interactivity between the viewer and the data visualization. The respondents can navigate the data visualization by hovering over the chart and viewing the information of that specific chart. They also can select the data by filtering data from the slicers.

Most respondents agreed that the display in mobile view is quite inconvenience and difficult to navigate compared to the desktop view. Overall, the respondents were satisfied viewing the exploratory dashboard as they enjoy the navigation from one chart to another chart to visualize the alumni data.

### E. Challenges

The greatest concern in this study is the privacy issue related to alumni personal information such as home addresses, grade point, and working place. The original data set need to be pre-processed before it can be used to develop the exploratory dashboard to ensure that the private information of the alumni is eliminated from the study following the ethics of research. This can be seen as a limitation since the alumni data visualization dashboard are not able to display more variety of data visualization types such as temporal data visualization, network data visualization, and geospatial data visualization.

## V. CONCLUSION

The importance of this study is to use data visualization in presenting alumni data through an exploratory dashboard and visualization techniques. The dashboard development process consists of three main phases, which are the conception phase, visualization phase and finalization. The exploratory dashboard is developed to visualize the alumni data using multidimensional and hierarchical data visualization techniques. The multidimensional data visualization includes a pie chart to visualize the gender of the alumni and a horizontal and vertical bar chart to visualize the age range, program name, and salary range of the alumni. The hierarchical data visualization includes a tree map chart to visualize the job domain of the alumni.

This alumni data visualization included a filtering function by embedding slicers in the exploratory dashboard to enable visitors to interact with the alumni data. The evaluation uses a convenience sampling method. The majority of the respondents expressed their satisfaction in evaluating the dashboard as they enjoy the fun of navigating from one chart to another chart to observe the alumni data. This dashboard can be used by the university's administrator to identify the gap between the qualification and salary range, and the relativeness of the job domain.

For future study, it is recommended to include other data fields such as state and alumni working place to have geospatial data visualization, and alumni networking data to be able to visualize network data visualization. It is recommended to have other alumni data that can support the other chart of types of data visualization such as the scatter plot, sunburst diagram, node-link diagram, and cartogram so that the alumni data visualization exploratory dashboard can be more informative and insightful.

#### ACKNOWLEDGMENT

The authors express gratitude to Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA for supporting this publication. The authors acknowledge the Ministry of Higher Education (MOHE) for funding under the Fundamental Research Grant Scheme (FRGS) (FRGS/1/2018/ICT04/UITM/02/9) and 600-IRMS/FRGS 5/3 (209/2019).

#### REFERENCES

- [1] A. A, "Challenges of alumni associations at universities: Income from alumni (donations and bequests) at South African universities", *African Journal of Business Management*, vol. 6, no. 45, pp. 11273-11280, 2012. Available: 10.5897/ajbm12.429. J. Clerk Maxwell, *A Treatise on Electricity and Magnetism*, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.
- [2] D. Dai, and Y. Lan, "The Alumni Information Management Model Based on "Internet +"", *Advances in Computer Science Research (ACSR)*, vol. 76, 2017.
- [3] K. Brush, and E. Burns, "What is data visualization and why is it important?", *SearchBusinessAnalytics*, 2020. [Online]. Available: <https://searchbusinessanalytics.techtarget.com/definition/data-visualization>
- [4] M. Aparicio, and C. Costa, "Data visualization", *Communication Design Quarterly*, vol. 3, no. 1, pp. 7-11, 2015. Available: 10.1145/2721882.2721883
- [5] E. Hayward, "The Starter Guide to Data Visualizations", *Klipfolio.com*, 2021. [Online]. Available: <https://www.klipfolio.com/resources/articles/what-is-data-visualization>
- [6] G. Wills, "Visualizing Hierarchical Data", *Encyclopedia of Database Systems*, pp. 3425-3432, 2009. Available: 10.1007/978-0-387-39940-9\_1380
- [7] "Hierarchy: The Data Visualisation Catalogue", *Datavizcatalogue.com*, 2022. [Online]. Available: <https://datavizcatalogue.com/search/hierarchy.html>
- [8] G. Wills, "Visualizing Network Data", *Encyclopedia of Database Systems*, pp. 3432-3437, 2009. Available: 10.1007/978-0-387-39940-9\_1381
- [9] "Network visualization: an introduction to visual network analysis", *Cambridge Intelligence*, 2022. [Online]. Available: <https://cambridge-intelligence.com/keylines/why-visualize-networks/>
- [10] "Effective Visualization of Multi-Dimensional Data—A Hands-on Approach", *Medium*, 2022. [Online]. Available: <https://medium.com/swlh/effective-visualization-of-multi-dimensional-data-a-hands-on-approach-b48f36a56ee8>
- [11] B. Soltoff, "Introduction to geospatial visualization | Computing for the Social Sciences", *Computing for the Social Sciences*, 2022. [Online]. Available: <https://cfss.uchicago.edu/notes/intro-geospatial-viz/>
- [12] K. Ebert, L. Axelsson and J. Harbor, "Opportunities and challenges for building alumni networks in Sweden: a case study of Stockholm University", *Journal of Higher Education Policy and Management*, vol. 37, no. 2, pp. 252-262, 2015. Available: 10.1080/1360080x.2015.1019117
- [13] "The Institute of Fine Arts, NYU," *Mapping The Institute of Fine Arts Alumni*. [Online]. Available: <https://ifa.nyu.edu/mapping-alumni/>
- [14] "UNIMAS Main Website", *Alumni* [Online]. Available: <http://www.alumni.unimas.my/what-we-offer>.
- [15] "Home", *Alumni Portal*, 2019. [Online]. Available: <https://alumni.iium.edu.my/>
- [16] "Alumni Summary Data | Ohio University", *Ohio.edu*, 2020. [Online]. Available: <https://www.ohio.edu/medicine/about/offices/assessment-and-accreditation/heritage-college-alumni-summary>
- [17] "Dashboard Development Process", *Unilytics*, 2016. [Online]. Available: <https://unilytics.com/data-visualization/dashboard-development-process/>
- [18] Salas-Velasco, M. Mapping the (mis)match of university degrees in the graduate labor market. *J Labour Market Res* 55, 14 (2021). <https://doi.org/10.1186/s12651-021-00297-x>
- [19] C. Cheng, "Experts: Fresh grads' pay as low as RM1000 a systematic problem, , can't just blame Covid-19 pandemic," *ISIS*, 15 April 2021. [Online]. Available: <https://www.isis.org.my/2021/04/15/experts-fresh-grads-pay-as-low-as-rm1000-a-systemic-problem-cant-just-blame-covid-19-pandemic/>

# The Performance Evaluation of Transfer Learning VGG16 Algorithm on Various Chest X-ray Imaging Datasets for COVID-19 Classification

Andi Sunyoto<sup>1\*</sup>, Yoga Pristyanto<sup>2</sup>, Arief Setyanto<sup>3</sup>

Fawaz Alarfaj<sup>4</sup>, Naif Almusallam<sup>5</sup>, Mohammed Alreshoodi<sup>6</sup>

Computer Science Department, Universitas Amikom Yogyakarta, Yogyakarta, Indonesia<sup>1, 2, 3</sup>

Computer & Information Sciences Department, Imam Mohammad Ibn Saud Islamic University, Kingdom of Saudi Arabia<sup>4, 5</sup>

Department of Natural Applied Science, Applied College, Qassim University, Buraydah, Kingdom of Saudi Arabia<sup>6</sup>

**Abstract**—Early detection of the coronavirus (COVID-19) disease is essential in order to contain the spread of the virus and provide effective treatment. Chest X-rays could be used to detect COVID-19 at an early stage. However, the pathological features of COVID-19 on chest X-rays closely resemble those caused by other viruses. The visual geometry group-16 (VGG16) deep learning algorithm based on convolutional neural network (CNN) architecture is commonly used to detect various pathologies on medical images automatically and may have a role in the detection of COVID-19 on chest X-rays. Therefore, this research is aimed to determine the robustness of the VGG16 architecture on several chest X-ray databases that vary in terms of size and the number of class labels. Nine publicly available chest X-ray datasets were used to train and test the algorithm. Each dataset had a different number of images, class compositions, and interclass proportions. The performance of the architecture was tested using several scenarios, including datasets above and below 5,000 samples, label class variation, and interclass ratio. This study confirmed that the VGG16 delivers robust performance on various datasets, achieving an accuracy of up to 97.99%. However, our findings also suggest that the accuracy of the VGG16 algorithm drops drastically in highly imbalanced datasets.

**Keywords**—Covid-19; Chest X-Ray; CNN; transfer learning; VGG-16

## I. INTRODUCTION

X-ray images are often used to detect changes in the lungs, such as pneumonia caused by viral infections. Pneumonia is also one of the key indicators of an infection caused by coronavirus disease (COVID-19). However, the manual evaluation of X-ray images is time-consuming and often subjective. Artificial intelligence (AI) could be used to automatically distinguish infected and infection-free patients by extracting specific shapes and spatial features visible on chest X-ray images. Many studies have been carried out using X-ray images to detect Middle East Respiratory Syndrome Coronavirus (MERS CoV) since there are features on chest X-rays and CT that resemble pneumonia manifestations [1]. A convolutional neural network (CNN) model has been developed to identify the nature of the pulmonary modulus on CT images and diagnose pneumonia on chest X-ray images [2].

COVID-19 symptoms include cough, fever, dyspnea, and respiratory problems. In more severe cases, COVID-19 can cause pneumonia, acute respiratory distress, septic shock, failure of internal organs, or even death [4]. Reverse-transcription-polymerase chain reaction (RT-PCR) of samples obtained from either blood or the respiratory system is often used to diagnose COVID-19. Furthermore, due to the highly infectious rate of COVID there is a high demand for this service which leads to further delays to obtain the test results. Therefore, in the emergency department, the initial diagnosis of symptomatic patients is more likely to be done through a plain chest X-ray or CT scan. The early identification of COVID-19 on plain X-rays or CT images is essential to isolate patients and hence minimize the spread of the disease as well as to treat infected patients more effectively.

Bilateral pulmonary parenchymal ground-glass and consolidative pulmonary opacities, with a rounded shape and a peripheral lung distribution, are common chest X-rays in COVID-19 patients [3]–[5]. Pneumonia is also an important indicator of COVID-19. However, these pathological features may closely resemble those caused by other viral infections, which makes it difficult for the radiologist to identify the type of infection. Deep learning algorithms based on convolutional neural network (CNN) architecture could be used to identify specific COVID-19 features on X-ray images.

CNN algorithms are easy to model and reliable. As a result, they are currently the most widely used artificial intelligence (AI) model for the detection of COVID-19 on X-ray images [6]–[13]. We reviewed several studies that made use of the CNN architecture to diagnose COVID-19 on chest X-rays, as shown in Table I. Our findings indicate that the CNN model developed by the visual geometry group with 16 depth layers (VGG16) has been applied in about 50% of the COVID-19 studies [6], [8], [9], [13]–[15]. The VGG16 also performed very well when compared with other established models. However, although studies based on the VGG16 models reported high levels of accuracy, research on COVID-19 is still evolving. Furthermore, most of the studies were based on a single dataset, potentially limiting the generalizability of the model. Therefore, the efficacy of the VGG16 model needs to be tested further on new emerging datasets. Because of these facts, we assessed additionally; we aimed to identify more

\*Corresponding Author.



pathological features on chest X-rays indicative of COVID-19 and other infectious diseases. This research contributes by testing the performance of VGG-16 on large, popular datasets and determining its level of accuracy. The results of this study will help to detect indications of COVID-19 more accurately in x-ray images and obtain alternative diagnoses of symptoms similar to those of COVID-19.

## II. METHODS

The research framework was divided into four steps: identification of publicly available chest X-ray datasets, image preprocessing, training and finally testing of the VGG16 algorithm, as shown in Fig. 1.

### A. Data Collection

Several publicly available image data repositories were searched using the keywords "COVID-19 X-ray". The search yielded 28 datasets composed of chest X-ray images used to diagnose COVID-19. The files within these datasets were reviewed to ensure that they contained chest X-ray imaging files that could be read by the COVID-19 detection system. Only the data set including big data i.e. containing more than 1000 images were included in the study. A total of nine datasets met the inclusion criteria. The final total datasets consisted of a total of 38,181 X-ray images. The images were classified into four categories: normal, pneumonia, viral pneumonia, and COVID-19, as shown in Fig. 2. However, the number of classifiers varied between the different datasets. The dataset source, the total number of images, and the number of classifiers within each dataset are summarized in Table II. Table II shows sample images for each class and the percentage number of images per class.

### B. Image Preprocessing

The VGG16 architecture uses a kernel size of 3x3 with an input image of 224x224x3 for the width, height, and channel, respectively [29]. In the preprocessing stage, the VGG16 input

was scaled to 224x224 pixels. The images of the classes were randomly extracted, and finally 80% were used to train the algorithm, 20% were used for validation, and the final 10% were used for testing as shown in Fig. 3.

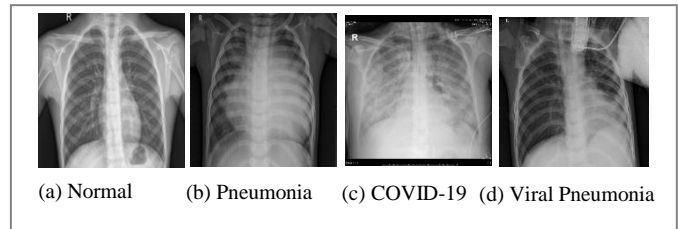


Fig. 1. Chest X-rays Illustrating the Four Different Classifiers used in the Datasets.

TABLE I. SUMMARY OF RESEARCH STUDIES COMPARING THE ROBUSTNESS OF SEVERAL CNN ARCHITECTURES FOR THE DETECTION OF COVID-19 ON CHEST X-RAYS

Author	Methods	Best Result
[6]	ResNet50, InceptionV3, and VGG16	VGG16
[7]	DenseNet-169+SVM, VGG16, RetinaNet + Mask RCNN, VGG16 and Xception	ResNet50
[8]	VGG16, VGG19, ResNet, DenseNet, and InceptionV3	VGG16
[9]	VGG16, MobileNetV2, Xception, NASNetMobile and InceptionResNetV2	VGG16
[10]	VGG16, ResNet50, and EfficientNetBo	EfficientNetBo
[16]	VGG16	VGG16
[12]	VGG16, DenseNet-161, ResNet-18	ResNet-18
[13]	MobileNet-V2 and VGG16	VGG16
[17]	AlexNet, VGG16, GoogleNet, MobileNet-V2, SqueezeNet, ResNet-34, ResNet-50 and Inception-V3	ResNet-34

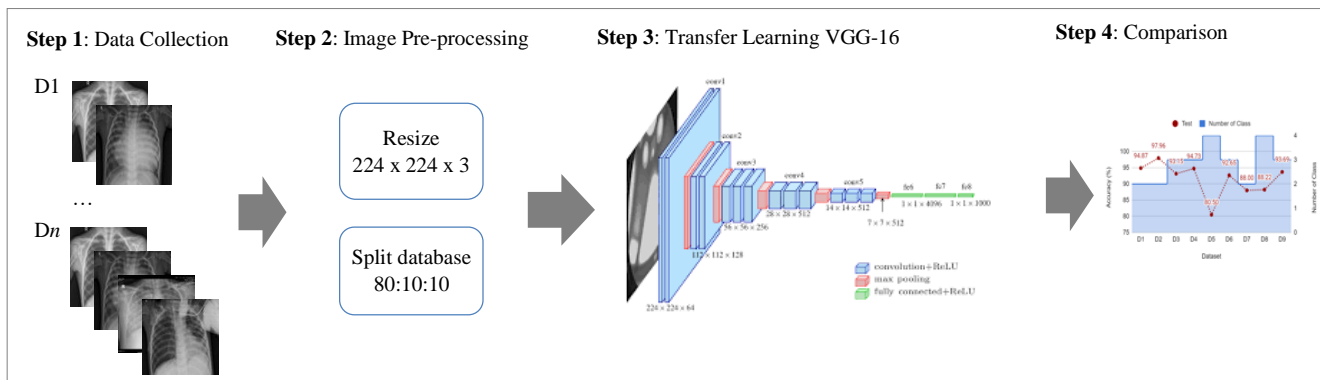


Fig. 2. Research Framework used to Conduct the Study.



TABLE II. DATASETS USED TO EVALUATE THE ROBUSTNESS OF THE VGG16 MODEL

Data	Name	Source	Classes	Files
D1	Chest X-ray Images (Pneumonia)	[18]	Normal:1583 Pneumonia : 4273	5,856
D2	CoronaHack -Chest X-ray-Dataset	[19], [20]	Normal:1576 Pneumonia : 4334	5,910
D3	COVID-19 Radiography Database	[21], [22]	COVID-19:219 Normal:1341 Viral Pneumonia : 1345	2,905
D4	Chest X-ray (COVID-19 & Pneumonia)	[23]– [25]	COVID-19:576 Normal:1583 Viral Pneumonia : 4273	6,432
D5	COVID-19 Detection X-ray Dataset	[18]– [20]	Bacterial Pneumonia:650 COVID-19:60 Normal:880 Viral Pneumonia : 412	2,002
D6	Covid-GAN and Covid-Net mini Chest X-ray	[18]– [20], [26]	COVID-19:461 Normal:1583 Pneumonia:4489	6,533
D7	COVID-19 X-ray Images5	[27]	COVID-19:60 Normal: 880	940
D8	Curated Chest X-ray Image Dataset for COVID-19	[25]	COVID-19:1281 Normal:3278 Pneumonia- Bacterial:3001 Pneumonia-Viral: 1656	6,515
D9	COVID-19 X-ray Dataset with Preprocessed Images	[19], [28]	COVID-19:361 Normal:365 Pneumonia:362	1,088

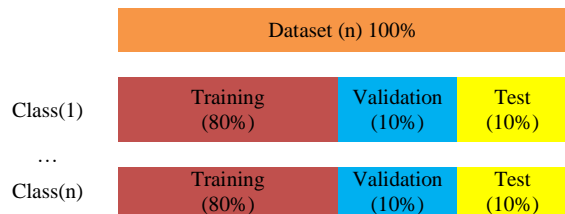


Fig. 3. Distribution of the Dataset for Training, Validation, and Testing.

### C. Application of the Convolutional Neural Network

CNN architectures are commonly used in computer vision and involve a convolution operation between the input signal and the filter. An important step in the development of the CNN algorithm involves the use of data pooling and convolution operations. During data pooling, the datasets are downsampled by averaging the data (average pooling) or by obtaining the maximum value (max pooling). In this case, the input signal features were derived from the chest X-ray image, which is commonly represented as  $n*m*c$ , whereby  $n$  and  $m$  represent the image width and length, respectively, and  $c$  represents the color channels. For example, in a typical red-green-blue (RGB) image with a pixel matrix size of  $256 \times 100$ , the input matrix would be defined as  $256 \times 100 \times 3$ . Convolution and pooling operations reduce the complexity by extracting only the important features. For example, an input signal consisting of 75,000 features can easily be reduced into 512 features by applying several convolution layers.

Several CNN-based architectures have been proposed in the last decade. Lenet-5 architecture was the first to propose a CNN-based architecture to solve a simple handwritten digit recognition problem [30]. The work was based on the older concepts of neural networks and back-propagation. The big leap of CNN-based algorithms was enabled by the availability of a large, labeled image dataset called Imagenet [31]. The dataset currently contains around 14 million labeled images, and it was initiated in 2009 by an artificial intelligence lab at Stanford University. Alexnet is the second most well-known CNN architecture that had won the Imagenet Large Scale Visual Recognition Challenge in 2014 [32].

The VGG16 model was initially proposed by Simonyan et al. [29] and it secured first place for object localization and second place for object classification in the Imagenet Large-Scale Visual Recognition Challenge 2014 (ILSVRC 2014). Since then, numerous other CNN architectures have been proposed including Inception net [33], ResNet [34], Inception-v4 and Inception-ResNet [33], Mobilenet [35], MobileNet V2 [36], EfficientNet [37] and XceptionNet [38].

The building block of the CNN architecture consists of two fundamental components: a convolution layer and a pooling layer. The filter size, padding, stride, activation function, and connection between layers can be manipulated to improve the performance of the algorithm. In order to improve the performance of the algorithm, the raw image signal has to be converted into a more straightforward representation before applying the classification task. The large number of images available on the Imagenet database could be used to train and compare the performance of various algorithms. Once an architecture has been trained and tested on the Imagenet images, researchers can ascertain the performance of their architectures and determine the weight parameters.

Transfer learning is a commonly used method in computer vision that applies the knowledge gained from the training of a network to solve a specific problem to a new similar scenario. This eventually reduces the time required for the training process allowing for the development of accurate algorithms in a shorter amount of time [29]. For example, if a pre-trained network previously developed to classify 1000 classes is now used for binary classification, the top layer (last layer) is adjusted so that the output is changed from 1000 into two classes only. The weight parameters also have to be updated in all network layers or for some of the layers. However, the learning process does not begin from scratch; instead, it starts with the pre-trained weight that has been applied to solve the previous problem.

Most model architectures such as VGG16, InceptionNet, mobilenet, and XceptionNet were trained on a large dataset such as Imagenet. As a result of the high computational cost incurred during training, an improved model was developed. Canziani et al. [39] conducted a comprehensive analysis of the performance of pre-trained models on computer vision challenges using data from the Imagenet [31] database. In computer vision, the transfer approach is popular because it enables the generation of accurate models in a shorter amount of time [40].

A typical CNN classification task has a feature generator and a classifier. The feature map generator input are the raw images, followed by a stack convolution and pooling layers. The main goal of the feature generator is to produce an array representing the input image in a smaller amount of data. On the other hand, the classifier's task is to categorize the feature into certain target classes. This task can be performed by classic classifiers such as support vector machines and decision trees. Another option is to place the fully connected layers on top of the feature generator. A fully connected layer is one whose neurons fully connect to all activations in the previous layer. The number of layers in the fully connected layer is significant and could be optimized by the researcher manually. The depth of the fully connected layers should be taken into account as it relates to the overfitting problem of the entire network. The deep learning approach independently computes the important input features during the learning process. Unlike the classical AI algorithm whereby the features extracted by the algorithm are based on the objective of the classification task and the image input, deep learning models learn hierarchical features by adjusting weight parameters on the CNN-based feature generator. The pattern of the input is then captured by the network and is then used as the input of the classifier. The pattern for a specific problem domain is accurately recorded as the weight parameter value. In transfer learning, the set of values can be applied to another specific problem domain.

Transfer learning involves two key steps. The first step involves selecting a pre-trained model, such as VGG16 [29], InceptionV3 [41], and ResNet5 [34], to fit the target problem. The second step involves the identification of features based on the size of the dataset and the similarities between the pre-trained dataset and the dataset we used. The comparison between the pre-trained dataset and our characteristics dataset could result in one of the following four transfer learning problems whereby the new problem dataset are:

- 1) large but dissimilar from the pre-trained dataset
- 2) large and highly similar to the pre-trained dataset
- 3) small and highly similar to the pre-trained dataset
- 4) small and dissimilar from the pre-trained dataset

In deep learning, a dataset consisting of 1000 labeled images per class is considered to be small. Dataset similarity refers to the availability of the same problem subset in the pre-trained dataset. For example, if the task is to recognize dogs and birds using a pre-trained network that has been trained on an Imagenet dataset since the dataset contains dog and bird classes, we can consider that the dataset is highly similar. However, in this paper, we classified the visual pathological features of COVID-19 patients through visible X-ray images which were not available in the Imagenet dataset. Hence, our transfer learning problem was due to a small dataset dissimilar from the pre-trained dataset.

This study employed the VGG16 [41], a pre-trained CNN-based architecture that consists of 16 CNN layers. The VGG architecture has already been applied in many medical image classification tasks [42]. This research provides an evaluation and comparison of the performance of various architectures in detecting COVID-19 on various chest X-ray datasets.

The comprehensive evaluation was made based on the accuracy achieved by the VGG16 architecture when applied to different datasets. Since the number of classes within the dataset varied, the multiclass confusion matrix was used to determine the robustness of the VGG16. Table III illustrates a confusion matrix with  $n$  classes, including the experimental results obtained for each scenario.

TABLE III. MULTICLASS CONFUSION MATRIX

		Predicted Number			
		Class 1	Class 2	...	Class n
Actual Number	Class 1	$x_{11}$	$x_{12}$	...	$x_{1n}$
	Class 2	$x_{21}$	$x_{22}$	...	$x_{2n}$
	...	...	...	...	...
	Class n	$x_{n1}$	$x_{n2}$	...	$x_{nn}$

Furthermore, the performance of the VGG16 model was quantified by calculating the total numbers of false-negative (TFN), false positive (TFP), true negative (TTN), total true positive (TTP) for each class  $i$  based on the general equations 1 to 4.

$$TFN_i = \sum_{j=1, j \neq i}^n x_{ij} \quad (1)$$

$$TFP_i = \sum_{j=1, j \neq i}^n x_{ji} \quad (2)$$

$$TTN_i = \sum_{j=1, j \neq i}^n \sum_{k=1, k \neq i}^n x_{jk} \quad (3)$$

$$TTP_i = \sum_{j=1}^n x_{ij} \quad (4)$$

In addition, the precision (P), recall (R), and specificity (S) for each class  $i$  were also calculated as shown in equations 5, 6 and 7. The total accuracy and F1-Score were calculated as shown in equations 8 and 9, respectively.

$$P_i = \frac{TTP_{all}}{TTP_{all} + TFP_i} \quad (5)$$

$$R_i = \frac{TTP_{all}}{TTP_{all} + TFN_i} \quad (6)$$

$$S_i = \frac{TTN_{all}}{TTN_{all} + TFP_i} \quad (7)$$

$$Accuracy = \frac{TTP_{all}}{Total\ Number\ of\ Testing} \quad (8)$$

$$F1 - Score = \frac{TTP_{all}}{Total\ Number\ of\ Testing} \quad (9)$$

### III. RESULT AND DISCUSSION

In this study, we made use of nine open databases to measure the performance of the CCN transfer learning model, VGG16, to detect pneumonia and COVID-19 cases. Each database contained more than 1000 images and a total of 38,181 images were evaluated. The datasets were divided according to the number of images and classifiers. The robustness of the VGG16 model was then tested on datasets that were.

- 1) with more than 5000 images.
- 2) with less than 5000 images.

3) with a different number of classes.

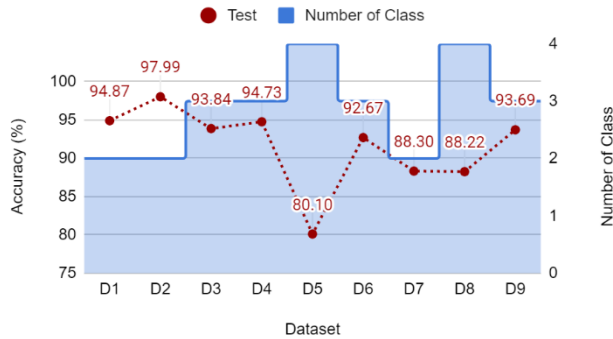


Fig. 4. Relationship between the Number of Classes in the Dataset and the Accuracy of the Algorithm.

Fig. 4 shows the relationship between the number of classes in the dataset and the resulting performance of the algorithm in terms of accuracy. The accuracy of the algorithm ranged from 80.10% to 97.99%. The algorithm performed better in datasets with three classifiers when compared with datasets with two classifiers. The algorithm had the worst performance in datasets with four classifiers.

The performance of the VGG16 algorithm was validated by splitting the dataset into three categories; training (80%), validation (10%), and test (10%). Fig. 5 illustrates the results of the confusion matrix from data test, while Table IV compares the accuracy, precision, recall, and F1-score of the VGG16 model in all of the nine databases evaluated in this study.

The testing accuracy analysis of the VGG16 algorithm in datasets containing more than 5,000 images is illustrated in Fig. 6. For this evaluation, we compared the accuracy of the VGG16 algorithm in D1, D2, D4, D6, and D8 datasets, which contained 5,856, 5,910, 6,432, 6,533, and 6,515 images, respectively. Following testing, the VGG16 algorithm in datasets with more than 5,000 images ranged from 97.99% to 88.22%. The findings of this analysis indicate that for both the validation and testing the detection accuracy of the algorithm decreased as the number of images increased.

The accuracy of the VGG16 algorithm on datasets containing less than 5000 images is illustrated in Fig. 7. For this analysis, the D7, D9, D5, and D3 datasets were used, which consisted of 940, 1,088, 2,002, and 2,905 images, respectively. For the testing data, the mean accuracy of the VGG16 algorithm was 94.31% with a range of 80.10-93.84%. As shown in Fig. 7, the difference in accuracy within each dataset was relatively small except for the D5 dataset, which showed an accuracy of 80.10%. However, further analysis showed that the low level of accuracy in the D5 dataset was caused by the very high-class difference ratio within this dataset.

Fig. 8 illustrates the performance of the VGG16 algorithm on datasets with two, three, and four classes. As evident in Fig. 8, the datasets with two and three classes did not differ much in terms of accuracy, but when a dataset has four classes, the accuracy decreases by 9.57%.

Based on the results of this study, we can conclude that the performance of the VGG16 algorithm is affected by the number of images and the number of classes within the dataset: For datasets with more than 5000 images, the accuracy of the algorithm decreased as the number of images in the dataset increased. The VGG16 model achieved a mean accuracy of 93.7%. Compared with previous studies, the VGG16 algorithm performed well despite its relatively simple architecture.

For datasets containing less than 5000 images, the number of images did not impact the algorithm's accuracy except for the D5 dataset. However, the lower performance of the algorithm in D5 was attributed to the larger class ratio within this dataset. The number of classes within a dataset affected the accuracy of the algorithm, whereby the VGG16 model performed worse in datasets with four classes. However, further research is recommended to test the efficacy of the transfer learning VGG16 model on the detection accuracy of COVID-19.

Compared to other popular transfer learning, the advantage of the VGG16 architecture is that it has only six layers in depth. The small layer makes the identification process fast. This fast time allows it to be applied to devices that have low specifications and are mobile. In real conditions, the dataset is not ideal, with different numbers and ratios between different classes. Based on the experiment, VGG16 solved cases of COVID-19 data with the characteristics of having a small class due to imbalanced data conditions. The limitation of VGG16 occurs in the unbalanced condition dataset, which has a large gap ratio (database D5 in Table II), and the number of classes is greater than four this show in Fig. 8.

TABLE IV. PERFORMANCE OF THE VGG16 ALGORITHM FOR ALL THE NINE DATABASES EVALUATED IN THE STUDY

Data	Name	Acc.	Prec.	Rec.	F1-Score
D1	Chest X-ray Images (Pneumonia) [18]	94.87	94.92	94.72	94.82
D2	CoronaHack -Chest X-ray-Dataset [19], [20]	97.99	97.65	97.29	97.47
D3	COVID-19 Radiography Database [21], [22]	93.84	90.48	88.68	89.57
D4	Chest X-ray (COVID-19 & Pneumonia) [23]–[25]	94.73	95.17	89.81	92.41
D5	COVID-19 Detection X-ray Dataset [18]–[20]	80.10	80.78	79.91	80.34
D6	Covid-GAN and Covid-Net mini Chest X-ray [18]–[20], [26]	92.67	89.43	85.75	87.55
D7	COVID-19 X-ray Images [27]	88.30	85.98	66.03	74.7
D8	Curated Chest X-ray Image Dataset for COVID-19 [25]	88.22	88.68	86.45	87.55
D9	COVID-19 X-ray Dataset With Preprocessed Images [19], [28]	93.69	93.69	94.03	93.86

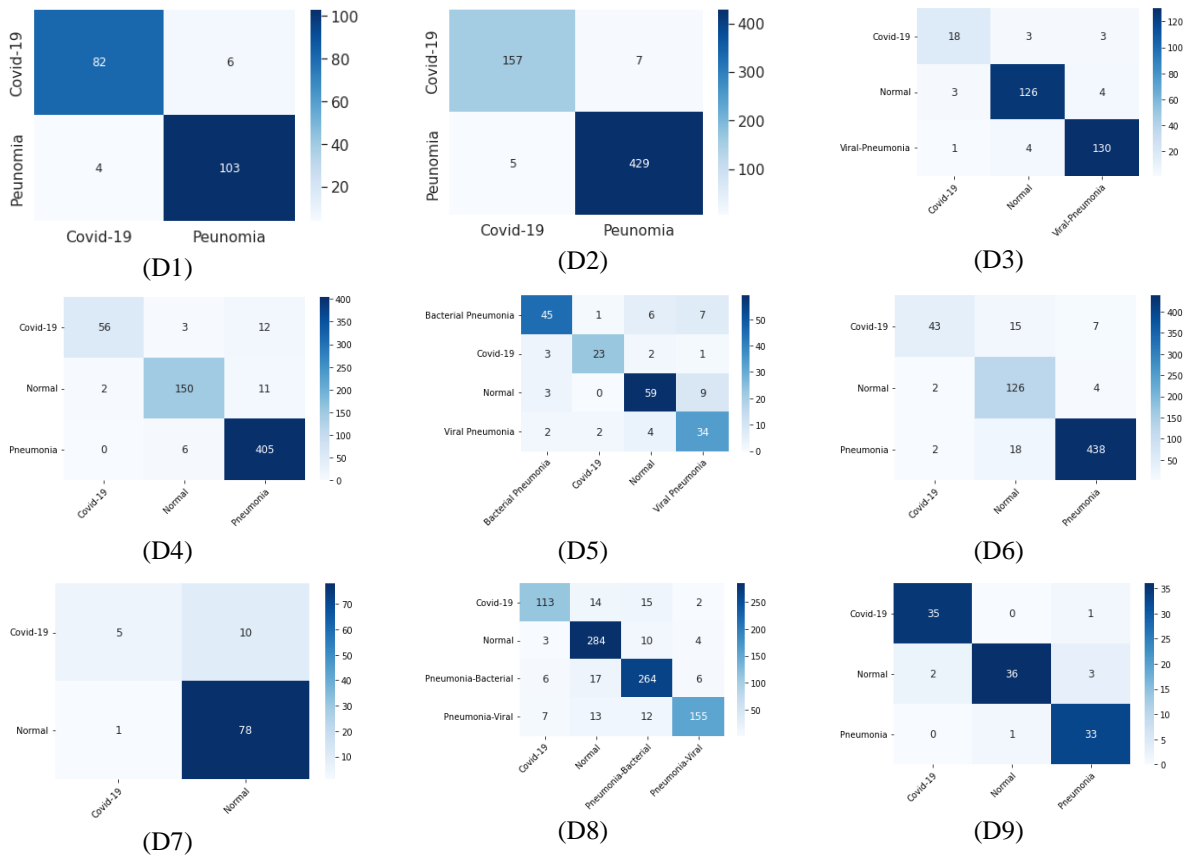


Fig. 5. Confusion Matrix Illustrating the Robustness of the VGG16 Algorithm in Different Datasets.

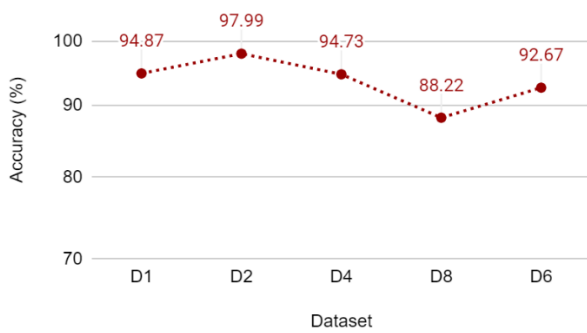


Fig. 6. Accuracy from Data Test of the VGG16 Algorithm in Datasets Containing more than 5,000 Images.

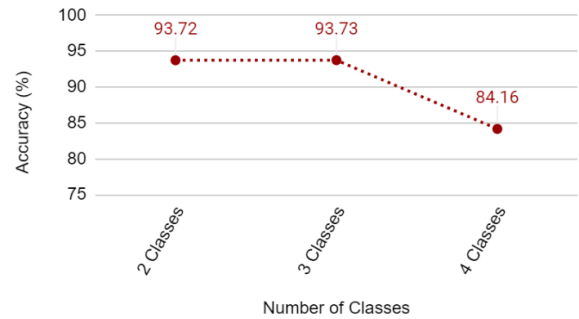


Fig. 8. Performance of the VGG16 Algorithm based on the Number of Classes.

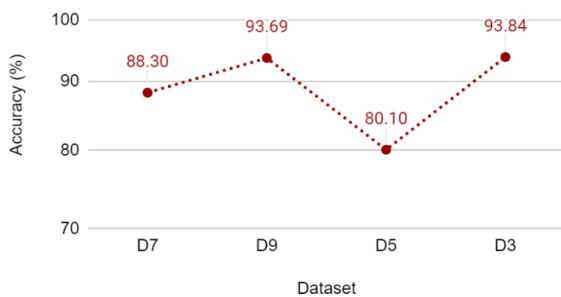


Fig. 7. Accuracy from Data Test of the VGG16 Algorithm in Datasets Containing Less than 5,000 Images.

#### IV. CONCLUSION

The aim of this study was to assess the performance of the VGG16 algorithm on different datasets. The experimental results confirmed the high accuracy of the VGG16 algorithm in detecting COVID-19. The study also confirmed the robustness of the VGG16 architecture when applied to datasets with various image numbers, classes, and class ratios on chest X-rays. However, in this study, we did not evaluate the impact of high-class ratios on the performance of the VGG16 algorithm. However, the class imbalance problem can easily be resolved via the application of data augmentation and class balancing techniques.

#### ACKNOWLEDGMENT

We would like to thank TopEdit ([www.topeditsci.com](http://www.topeditsci.com)) for its linguistic assistance during the preparation of this manuscript.

Funding statement: This research was funded by the Deanship of Scientific Research at Imam Mohammad Ibn Saud Islamic University through Research Group no. RG-21-51-01.

Conflicts of interest: The authors declare that they have no conflicts of interest to report regarding the present study.

#### REFERENCES

- [1] A. Hamimi, "MERS-CoV: Middle East respiratory syndrome corona virus: Can radiology be of help? Initial single center experience," *The Egyptian Journal of Radiology and Nuclear Medicine*, vol. 47, no. 1, pp. 95–106, Mar. 2016, doi: 10.1016/j.ejrm.2015.11.004.
- [2] J. Choe, S. M. Lee, K.-H. Do, G. Lee, J.-G. Lee, *et al.*, "Deep Learning-based Image Conversion of CT Reconstruction Kernels Improves Radiomics Reproducibility for Pulmonary Nodules or Masses," *Radiology*, vol. 292, no. 2, pp. 365–373, Aug. 2019, doi: 10.1148/radiol.2019181960.
- [3] D. Wang, B. Hu, C. Hu, F. Zhu, X. Liu, *et al.*, "Clinical Characteristics of 138 Hospitalized Patients With 2019 Novel Coronavirus-Infected Pneumonia in Wuhan, China," *JAMA*, vol. 323, no. 11, p. 1061, Mar. 2020, doi: 10.1001/jama.2020.1585.
- [4] C. Huang, Y. Wang, X. Li, L. Ren, J. Zhao, *et al.*, "Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China," *The Lancet*, vol. 395, no. 10223, pp. 497–506, Feb. 2020, doi: 10.1016/S0140-6736(20)30183-5.
- [5] M. Chung, A. Bernheim, X. Mei, N. Zhang, M. Huang, *et al.*, "CT Imaging Features of 2019 Novel Coronavirus (2019-nCoV)," *Radiology*, vol. 295, no. 1, pp. 202–207, Apr. 2020, doi: 10.1148/radiol.202002030.
- [6] S. Guefrechi, M. Ben Jabra, A. Ammar, A. Koubaa, and H. Hamam, "Deep learning based detection of COVID-19 from chest X-ray images," *Multimedia Tools and Applications*, Jul. 2021, doi: 10.1007/s11042-021-11192-5.
- [7] A. Manickam, J. Jiang, Y. Zhou, A. Sagar, R. Soundrapandian, *et al.*, "Automated pneumonia detection on chest X-ray images: A deep learning approach with different optimizers and transfer learning architectures," *Measurement*, vol. 184, p. 109953, Nov. 2021, doi: 10.1016/j.measurement.2021.109953.
- [8] K. Sahinbas and F. O. Catak, "Transfer learning-based convolutional neural network for COVID-19 detection with X-ray images," in *Data Science for COVID-19*, Elsevier, 2021, pp. 451–466. doi: 10.1016/B978-0-12-824536-1.00003-4.
- [9] A. K. Rangarajan and H. K. Ramachandran, "A preliminary analysis of AI based smartphone application for diagnosis of COVID-19 using chest X-ray images," *Expert Systems with Applications*, vol. 183, p. 115401, Nov. 2021, doi: 10.1016/j.eswa.2021.115401.
- [10] T. Zebin and S. Rezvy, "COVID-19 detection and disease progression visualization: Deep learning on chest X-rays for classification and coarse localization," *Applied Intelligence*, vol. 51, no. 2, pp. 1010–1021, Feb. 2021, doi: 10.1007/s10489-020-01867-1.
- [11] R. K. Singh, R. Pandey, and R. N. Babu, "COVIDScreen: explainable deep learning framework for differential diagnosis of COVID-19 using chest X-rays," *Neural Computing and Applications*, Jan. 2021, doi: 10.1007/s00521-020-05636-6.
- [12] A. Shelke, M. Inamdar, V. Shah, A. Tiwari, A. Hussain, *et al.*, "Chest X-ray Classification Using Deep Learning for Automated COVID-19 Screening," *SN Computer Science*, vol. 2, no. 4, p. 300, Jul. 2021, doi: 10.1007/s42979-021-00695-5.
- [13] H. Swapnarekha, H. S. Behera, D. Roy, S. Das, and J. Nayak, "Competitive Deep Learning Methods for COVID-19 Detection using X-ray Images," *Journal of The Institution of Engineers (India): Series B*, Apr. 2021, doi: 10.1007/s40031-021-00589-3.
- [14] R. M. James and A. Sunyoto, "Detection Of CT - Scan Lungs COVID-19 Image Using Convolutional Neural Network And CLAHE," in *2020 3rd International Conference on Information and Communications Technology (ICOIACT)*, Nov. 2020, pp. 302–307. doi: 10.1109/ICOIACT50329.2020.9332069.
- [15] N. Narayan Das, N. Kumar, M. Kaur, V. Kumar, and D. Singh, "Automated Deep Transfer Learning-Based Approach for Detection of COVID-19 Infection in Chest X-rays," *IRBM*, Jul. 2020, doi: 10.1016/j.irbm.2020.07.001.
- [16] M. Singh, S. Bansal, S. Ahuja, R. K. Dubey, B. K. Panigrahi, *et al.*, "Transfer learning-based ensemble support vector machine model for automated COVID-19 detection using lung computerized tomography scan data," *Medical & Biological Engineering & Computing*, vol. 59, no. 4, pp. 825–839, Apr. 2021, doi: 10.1007/s11517-020-02299-2.
- [17] S. R. Nayak, D. R. Nayak, U. Sinha, V. Arora, and R. B. Pachori, "Application of deep learning techniques for detection of COVID-19 cases using chest X-ray images: A comprehensive study," *Biomedical Signal Processing and Control*, vol. 64, p. 102365, Feb. 2021, doi: 10.1016/j.bspc.2020.102365.
- [18] D. S. Kermany, M. Goldbaum, W. Cai, C. C. S. Valentim, H. Liang, *et al.*, "Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning," *Cell*, vol. 172, no. 5, pp. 1122–1131.e9, Feb. 2018, doi: 10.1016/j.cell.2018.02.010.
- [19] J. P. Cohen, P. Morrison, L. Dao, K. Roth, T. Q. Duong, *et al.*, "COVID-19 Image Data Collection: Prospective Predictions Are the Future," Jun. 2020.
- [20] J. P. Cohen, P. Morrison, and L. Dao, "COVID-19 Image Data Collection," Mar. 2020.
- [21] M. E. H. Chowdhury, T. Rahman, A. Khandakar, R. Mazhar, M. A. Kadir, *et al.*, "Can AI Help in Screening Viral and COVID-19 Pneumonia?," *IEEE Access*, vol. 8, pp. 132665–132676, 2020, doi: 10.1109/ACCESS.2020.3010287.
- [22] T. Rahman, A. Khandakar, Y. Qiblawey, A. Tahir, S. Kiranyaz, *et al.*, "Exploring the effect of image enhancement techniques on COVID-19 detection using chest X-ray images," *Computers in Biology and Medicine*, vol. 132, p. 104319, May 2021, doi: 10.1016/j.compbiomed.2021.104319.
- [23] G. Maguolo and L. Nanni, "A Critic Evaluation of Methods for COVID-19 Automatic Detection from X-Ray Images," Apr. 2020, [Online]. Available: <http://arxiv.org/abs/2004.12823>.
- [24] E. Tartaglione, C. A. Barbano, C. Berzovini, M. Calandri, and M. Grangotto, "Unveiling COVID-19 from CHEST X-Ray with Deep Learning: A Hurdles Race with Small Data," *International Journal of Environmental Research and Public Health*, vol. 17, no. 18, p. 6933, Sep. 2020, doi: 10.3390/ijerph17186933.
- [25] U. SAIT, L. k VGokul, T. Prajapati, SunnyBhaumik RahulKumar, and K. SanjanaBHalla, "Curated Dataset for COVID-19 Posterior-Anterior Chest Radiography Images (X-Rays)," p. SAIT, UNAIS; k v, Gokul Lal; Prajapati, Sunny; Bha, doi: 10.17632/9xkhgts2s6.1.
- [26] C. D. Corporation, "Actualmed COVID-19 Chest X-ray Dataset Initiative," *Canada and Vision and Image Processing Research Group, University of Waterloo*, 2020. <https://github.com/agchung/Actualmed-COVID-chestxray-dataset>.
- [27] D. Uddipta, "COVID-19 X-ray Images5," <https://www.kaggle.com/uddiptadas/covid19-xray-images5>.
- [28] A. Gupta, Anjum, S. Gupta, and R. Katarya, "InstaCovNet-19: A deep learning classification model for the detection of COVID-19 patients using Chest X-ray," *Applied Soft Computing*, vol. 99, p. 106859, Feb. 2021, doi: 10.1016/j.asoc.2020.106859.
- [29] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," Sep. 2014.
- [30] L. Cun, L. Cun, B. Boser, J. S. Denker, D. Henderson, *et al.*, "Handwritten Digit Recognition with a Back-Propagation Network," *ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS*, vol. 2, pp. 396–404, 1990, Accessed: Dec. 01, 2021. [Online]. Available: <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.32.5076>.
- [31] J. Deng, W. Dong, R. Socher, L.-J. Li, Kai Li, *et al.*, "ImageNet: A large-scale hierarchical image database," in *2009 IEEE Conference on*

- Computer Vision and Pattern Recognition*, Jun. 2009, pp. 248–255. doi: 10.1109/CVPR.2009.5206848.
- [32] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” *Commun ACM*, vol. 60, no. 6, pp. 84–90, May 2017, doi: 10.1145/3065386.
- [33] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, “Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning,” Feb. 2016.
- [34] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, Dec. 2015, doi: 10.1109/CVPR.2016.90.
- [35] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, *et al.*, “MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications,” Apr. 2017.
- [36] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, “MobileNetV2: Inverted Residuals and Linear Bottlenecks,” in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2018, pp. 4510–4520. doi: 10.1109/CVPR.2018.00474.
- [37] M. Tan and Q. v Le, “EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks.”
- [38] F. Chollet, “Xception: Deep Learning with Depthwise Separable Convolutions,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul. 2017, pp. 1800–1807. doi: 10.1109/CVPR.2017.195.
- [39] A. Canziani, A. Paszke, and E. Cukurciello, “An Analysis of Deep Neural Network Models for Practical Applications,” May 2016.
- [40] [40] W. Rawat and Z. Wang, “Deep Convolutional Neural Networks for Image Classification: A Comprehensive Review,” *Neural Computation*, vol. 29, no. 9, pp. 2352–2449, Sep. 2017, doi: 10.1162/neco\_a\_00990.
- [41] S. Ioffe and C. Szegedy, “Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift,” *ArXiv*, vol. abs/1502.0, 2015.
- [42] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, *et al.*, “A survey on deep learning in medical image analysis,” *Medical Image Analysis*, vol. 42, pp. 60–88, Dec. 2017, doi: 10.1016/j.media.2017.07.005.



# A Comprehensive Review and Application of Interpretable Deep Learning Model for ADR Prediction

Shiksha Alok Dubey<sup>1</sup>, Anala A. Pandit<sup>2</sup>

Department of Computer Application  
Veermata Jijabai Technological Institute (V.J.T.I)  
Mumbai, India

**Abstract**—Drug safety is a pressing need in today's healthcare. Minimizing drug toxicity and improving the individual's health and society is the key objective of the healthcare domain. Drugs are clinically tested in laboratories before marketing as medicines. However, the unintended and harmful effects of drugs are called Adverse Drug Reactions (ADRs). The impact of ADRs can range from mild discomfort to more severe health hazards leading to hospitalization and in some cases death. Therefore, the objective of this research paper is to design a framework based on which research papers are collected from both ADR detection and prediction domain. Around 172 research articles are collected from the sites like ResearchGate, PubMed, etc. After applying the elimination criteria the author categorized them into ADR detection and prediction themes. Further, common data sources and algorithms as well as the evaluation metrics were analyzed and their contribution to their respective domains is stated in terms of percentages. A deep learning framework is also designed and implemented based on the research gaps identified in the existing ADR detection and prediction models. The performance of the deep learning model with two hidden layers was found to be optimum for ADR prediction and further, the non-interpretability part of the model is addressed using a global surrogate model. The proposed architecture has successfully addressed multiple limitations of existing models and also highlights the importance of early detection & prediction of adverse drug reactions in the healthcare industry.

**Keywords**—Drug safety; adverse drug reactions; early detection; deep learning; interpretable models

## I. INTRODUCTION

Drug safety is a pressing need of today's healthcare. Minimizing drug toxicity and improving the health of individuals and society is the key objective of the healthcare domain. The drug development process starting from discovery to market is long and costly. Rigorous efforts are involved in clinical trials to ensure the safety and efficacy of the developed drugs. Clinical trials are performed on any new drug substance to check its safety and effectiveness against the particular disease before marketing them as medicines to the general population [1]. Currently, all developed drugs have risks associated with them [2] and only those drugs whose curative impact is greater than the risk, are marketed as medicines. Any unwanted, undesired effects of drugs on human health are considered Adverse Drug Reactions

(ADRs). According to the definition provided by WHO (World Health Organization), an ADR can be unintended and occurs at doses normally used in man for the prophylaxis, diagnosis or therapy of disease, or modification of physiological function" [3]. Simply, it can be seen as an unpleasant or unexpected effect of drugs on patients. The impact of ADRs is considered to be one of the reasons behind mortality and morbidity in humans. The contribution of ADRs is about 5% of all hospital admission and it is considered the fifth most common reason for mortality during hospitalizations [4]. The severity and harmfulness of the reported ADRs have caused a ban on many developed drugs. In about 20 years around 40 drugs are withdrawn from the drug market due to the severe reactions caused due to them [5]. Around 50% are banned from the US market [6] then Germany [5] and finally from the Europe drugs market. The most common occurring toxicities due to drugs are cardiotoxicity (32%, [13]), hepatotoxicity (20%, [8]) then death risk (10%, [4]), and finally risk of overdose (7%, [3]). A recent example of a Sibutramine (Meridia) drug got initial permission from FDA to be sold as an appetite suppressant, but in 2010 it was banned from the market as it caused an increase in heart disease and heart stroke risk in patients. The severity of ADRs can also be measured in terms of the burden of healthcare cost and length of hospital stay [7]. A variety of factors are also responsible for the development of ADRs in humans. These factors can be classified as patient-related, drug-related, and social environment-related [8]. Gender and age are critical patient-related parameters that need to be considered while assessing the impact of ADRs on individuals while drug dosage and drug-drug interaction are important drug-related factors [8]. Confounding factors like smoking and alcoholism are crucial in the development of ADRs [8]. For better patient safety and improving healthcare, it is important to not only predict an ADR on time but also detect it at an early stage.

The following diagram Fig. 1, shown illustrates that ADRs are included as part of ADE (Adverse Drug Event) which is again a subset of adverse events. But the fact that separates ADRs from ADE is that they are caused due to drug intake even at normal dosage.

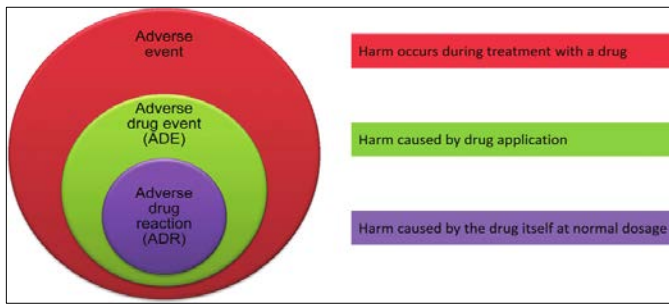


Fig. 1. Classification of Adverse Events. Adverse Events Include all Harmful Events Occurring during Treatment with a Drug without the Necessity of a Causal Link between the Drug and the Reaction. If the use of Medication is Causal to the Reaction, the Condition is called an Adverse Drug Event. A Subform of Adverse Drug Events is Adverse Drug Reactions that are Triggered by the Drug Itself Despite its Appropriate Dosage [9].

## II. FRAMEWORK FOR RESEARCH PAPER SELECTION

In the last two decades, a lot of research has been done in this field of ADR identification and improving drug safety. Researchers have published their research works highlighting the need for the detection and prediction of ADRs in the healthcare industry. Therefore the purpose is to first summarize these published research articles from multiple perspectives and then apply deep learning models for ADR prediction. The research papers are collected from both domains. Depending on the elimination criteria defined, only the relevant research works are selected for further analysis. Arksey and O'Malley's [10, 11] methodological framework is used for selecting research papers for literature review. This framework has helped researchers to concentrate on a single domain for a short duration and identify research gaps depending on the collected research works. The entire methodology can be summarized as follows:

### Stage 1: Identifying the research question

As previously discussed this review focuses mainly on the research studies done in the past related to the detection and accurate prediction of ADRs. Detecting an ADR from data is important before predicting it, therefore research papers are included from both domains.

#### Theme 1: ADR Detection

What makes ADR detection critical for drug safety? What is the different ADR signal detection techniques applied to datasets? How the different techniques are evaluated on a variety of datasets?

Detecting an ADR is an important step to improve healthcare and drug safety [12]. It is important to detect ADR and distinguish it from the symptoms of the disease. Different detection techniques are defined for different datasets.

#### Theme 2: ADR Prediction

Why accurate prediction of ADR is important for better patient safety and minimizing ADR occurrences? What is the different prediction models applied for ADR prediction? How computational models are useful in preventing severe ADRs in the future?

Predicting an ADR can prevent its occurrence and minimize healthcare costs [13]. Different models have been applied in the past, present, and future to predict and prevent such ADRs. The extent of this review study includes machine learning and deep learning models for ADR prediction [14].

### Stage 2: Collecting the research studies

As previously discussed, the author has collected research articles related to ADR detection and prediction domain published throughout 10yrs. The research studies are from both computer science and biomedical domain. Major search engines and databases from where these publications & databases used in those publications are:-

PubMed:-It is a search engine that provides easy access to the MEDLINE database [15] and is freely available. It also provides access to abstracts and references related to biomedical as well as life science domains [16].

ResearchGate:- It is a European social networking site [17] that provides a common platform for both researchers and scientists. The majority of research articles related to different domains are published on ResearchGate [18] for access to both researchers and academic professionals.

The indexing mechanism available in PubMed is Medical Subject Headings (MeSH) [19].

MeSH:- It is a controlled comprehensive vocabulary [19] used for indexing journals available on PubMed. This indexing is very helpful for searching research articles and journal papers. The research studies were searched using different keywords related to 'adverse drug reaction', 'prediction related ADR', and 'detection of ADR', and different datasets were openly accessible and acquired through ethical permission.

Query-based search: - The different query strings related to pharmacovigilance [20] are used for searching different articles on Google Scholar. The articles are searched based on heading, abstract and main content.

### Stage 3: Select only the relevant studies.

The author has defined some criteria based on which only the relevant research studies were selected. The elimination criteria are listed below:-

Duplicate research papers are eliminated.

Research studies not related to the review.

The research papers largely focused on the biomedical domain.

The research studies were more related to clinical research.

The research article more focused on drug-drug interaction and the genetic interaction of drugs.

Other unrelated research works.

After filtering, only relevant research works are selected for further analysis, and the results drawn are presented in this paper.

Stage 4: Charting the Data

The author has reviewed the research papers from multiple aspects. The various perspectives based on which the research studies are evaluated are described below:-

- Search Engine/Database
- Year of Publication
- Journal/Conference
- Name of the research paper
- Datasets used
- Models applied for ADR detection & prediction
- Drugs mentioned for a given ADR
- Evaluation metrics applied

Stage 5: Summarizing and reporting results

The research studies are summarized and segregated based on the approach used. In the initial phase, only the relevant studies are considered and the irrelevant ones are filtered out. Then the research works are grouped according to themes 1 and 2. Theme 1 is ADR detection whereas theme 2 is ADR prediction. The layout of the entire process is outlined in the flowchart shown in Fig. 2.

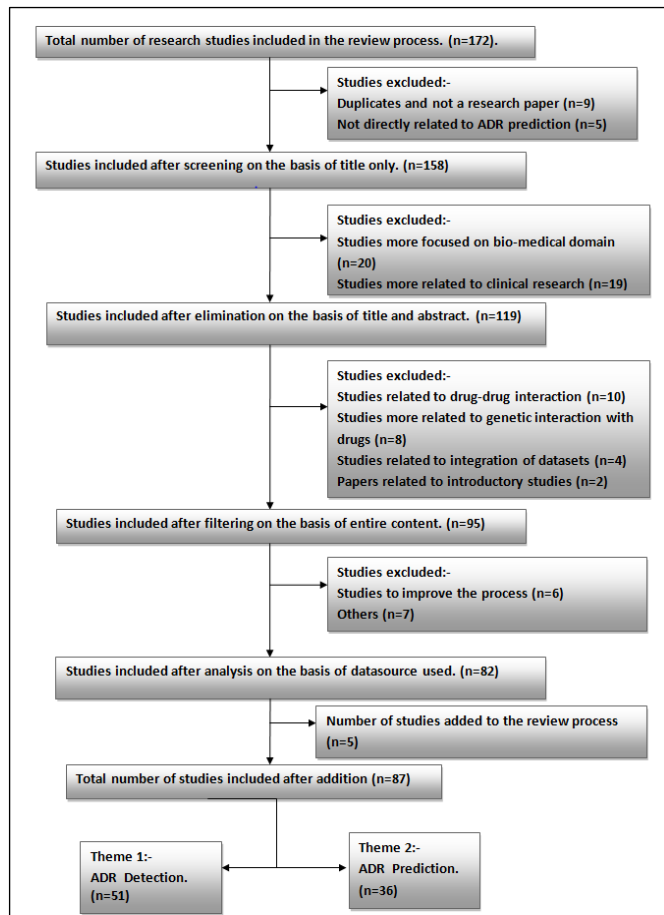


Fig. 2. Framework for Research Paper Selection.

The layout of the entire process is illustrated in the above diagram. Around 172 research works are collected showing the adverse effects and reactions of drugs on human health. The elimination criteria are used to eliminate irrelevant research papers. Repeated research works (n=9) and studies not related to ADR prediction (n=5) are filtered. Around 158 research works remained after elimination. About 19 research studies were eliminated in the screening process that belongs to the clinical research domain. Around 20 research works, more related to the biomedical domain were also eliminated. After filtering it only 119 research works remained. Research related to drug-drug interaction (n=10) and genetic interaction of drugs (n=8) were also screened out. Introductory studies (n=2) and research works related to the integration of datasets (n=4) are eliminated. A total of 95 research papers remained. Some research studies related to the improvement of ADR detection and prediction process (n=6) along with others (n=7) are also eliminated. In addition, five more research papers were made for the final review analysis. Finally, after filtering the research studies based on elimination criteria previously defined, the author identified about 87 research papers for further analysis.

Classifying them according to the two themes of ADR detection and prediction, there are about 51 papers associated with ADR detection and the remaining 36 are related to ADR prediction.

III. SUMMARY OF ADR DATASETS

Incidents of adverse reactions have been in existence for more than two decades. Over the period many countries have established pharmacovigilance centers [21] for collecting the reported occurrences of ADRs from medical practitioners and healthcare workers. These centers contribute to the postmarketing surveillance of ADRs. Many secondary data sources have been established by collecting both prescription data and ADR information. ADRs are also monitored actively through clinical trials and identified in different structured and unstructured data sources [22]. Different ADR-related data sources are listed and discussed in Table I.

TABLE I. SUMMARY OF ADR DATASETS

ADR related datasources	Description	Website
<b>Primary databases</b>		
<b>Spontaneous reporting Systems(SRS)</b>		
FAERS[23] EMA[24] UMC[25]	The incidents of ADRs are reported to the regulatory bodies of the country. They are analyzed and stored in databases for further action against the reported drug. These databases are also available for review and research process.	<a href="https://open.fda.gov/data/faers/">https://open.fda.gov/data/faers/</a> <a href="https://www.ema.europa.eu/en">https://www.ema.europa.eu/en</a> <a href="https://www.who-umc.org/">https://www.who-umc.org/</a>
Electronic Health Records[26]	This database contains records of patients admitted to the hospital. The datasource is very accurate as it records all information about the patient's condition the disease and its recovery phases.	This database can only be required through ethical permission from the required regulatory authority
Clinical narratives	The narratives and discharge summaries are written by experienced healthcare professionals. It contains	This data again requires ethical permission for

	data about a patient, disease, prescription information, and the treatment given. This information is very accurate and precise.	using it as part of the research.
<b>Major secondary ADR databases</b>		
SIDER 4.1[27] (Medical Literature)	This dataset includes data of 1430 marketed medicines and their recorded 5868 ADRs. It also includes around 139756 drug-SE association pairs.	<a href="http://sideeffects.embl.de/">http://sideeffects.embl.de/</a>
OFF-SIDES[28]	Offsides is a database of drug side effects that were found, but are not listed on the official FDA label.	<a href="http://tatonettlab.org/resources/nsides/">http://tatonettlab.org/resources/nsides/</a>
TWOSIDES [28]	An online available dataset containing information about drug-drug interaction and side-effects due to drug-drug interactions.	<a href="http://tatonettlab.org/resources/nsides/">http://tatonettlab.org/resources/nsides/</a>
ADReCS [29]	A comprehensive ADR ontology database.	<a href="http://bioinf.xmu.edu.cn/ADReCS">http://bioinf.xmu.edu.cn/ADReCS</a>
Medical Forums	These are public websites used for posting health-related inquiries.	<a href="https://www.dailystrength.org/">https://www.dailystrength.org/</a>
<b>Major API(Active Pharmaceutical Ingredients) interaction databases</b>		
DrugBank [30]	Comprehensive online database containing information on drugs & drug targets.	<a href="https://go.drugbank.com/">https://go.drugbank.com/</a>
PubChem [31]	A resource with information on chemical substances and their biological activities	<a href="https://pubchem.ncbi.nlm.nih.gov/">https://pubchem.ncbi.nlm.nih.gov/</a>
SuperDRUG 2[32]	SuperDRUG2 is a unique, one-stop resource for marketed and approved drugs containing 4,600 active pharmaceutical ingredients [32].	<a href="http://cheminfo.charite.de/superdrug2/">http://cheminfo.charite.de/superdrug2/</a>
SuperTarget	A resource that contains information about drug and target proteins and analyses their associations.	<a href="https://bioinformatics.charite.de/supertarget/">https://bioinformatics.charite.de/supertarget/</a>
STITCH[33]	A resource collecting known and predicted interactions between chemicals and proteins.	<a href="http://stitch.embl.de/">http://stitch.embl.de/</a>
PharmGKB [34]	The pharmacogenomic knowledgebase is a publicly available online knowledge base used for aggregation and integration of information on drugs and analyzing their impact on genetic variation.	<a href="https://www.pharmgkb.org/">https://www.pharmgkb.org/</a>
KEGG(Kyoto Encyclopedia of Genes & Genomes ) [35] & GO(Gene Ontology) [36]	It is a collection of databases dealing with genomes, biological pathways, diseases, drugs, and chemical substances. GO resource contains information about gene function.	<a href="https://www.genome.jp/kegg/">https://www.genome.jp/kegg/</a> <a href="http://geneontology.org/">http://geneontology.org/</a>

Different related data sources are grouped into separate categories. The basic categories defined are primary and secondary data sources. The table incorporates a variety of ADR-related data resources for both ADR detection and prediction.

#### A. ADR Detection

The research studies are majorly done in USA (n=33/51, 65%), Europe (n=6/51, 12%) and Korea (33/51, 8%). Apart from this research contributions from Australia (n=3/51,6%) and India (n=2/51,4%) are also considered.

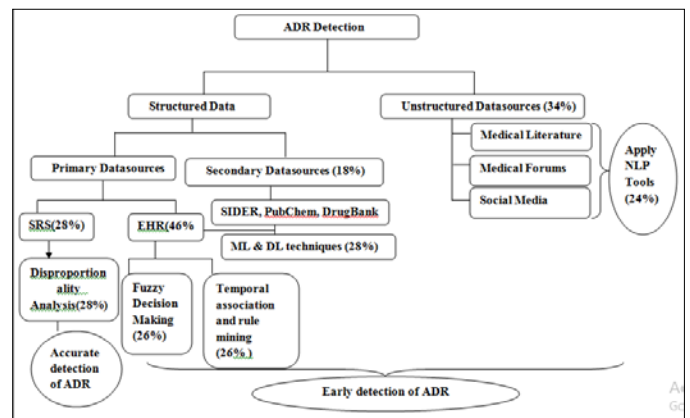


Fig. 3. The Layout of the ADR Detection Process.

As shown in Fig. 3, the ADR detection data sources are categorized into two groups that are structured and unstructured datasets. The unstructured dataset is used in about 34% (n=17/51) of the research works while structured datasets are again categorized into primary and secondary data sources. The primary data sources are again divided into SRS and EHR. SRS is utilized in around 28% (n=14/51) of the research papers while EHR is in 46% (n=23/51) of the research. The secondary data source includes information about the drug-ADR association and is included in about 18% (n=9/51) of the research works. The different techniques are applied based on the data sources used.

DPA (Disproportionality analysis) is applied in around 28% of research papers where SRS is involved to validate the potential drug-ADR association.

Fuzzy Decision Making & Temporal Association Mining is applied equally in about 26% of the research studies for the early detection of ADRs.

Machine Learning (ML) & Deep Learning (DL) models are applied for secondary and primary data sources in around 28% of the research studies. The models are trained to detect unknown drug-ADR associations from datasets.

Finally, NLP (Natural Language Processing Tools & Techniques) are applied in about 24% of the research studies for extracting meaningful insights from unstructured text.

EHR and unstructured text has been used to early detect an ADR while SRS is helpful in the accurate detection of ADR.

#### B. ADR Prediction

The geographical research distribution for ADR prediction shows that the majority of research work is carried out in the USA at 44 % ( n=16/36) followed by China at 22 % ( n=8/36) and finally in Europe at 17 % ( n=6/36). Other countries like Croatia, Romania, India, Israel, Iran, Korea, and Japan have also contributed to the research in the ADR prediction domain. The major steps performed for ADR prediction are illustrated as follows:-



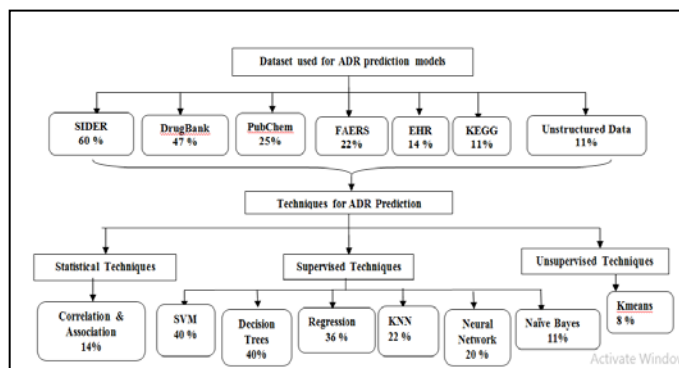


Fig. 4. The Layout of the ADR Prediction Process.

The Fig. 4, illustrates that the datasets mainly used for ADR prediction are SIDER 60% (n=21/36), Drugbank 47% (n=17/36), PubChem 25% (n=9/36) and FAERS 22% (n=8/36). Further, the techniques applied for ADR prediction are divided into three categories that are statistical, supervised, and unsupervised techniques. The statistical methods are further defined as correlation and association methods that contribute to 14% (n=5/36) of the research works while unsupervised techniques are further classified as Kmeans are applied in about 8% (n=3/36) of the research works. The common supervised techniques applied in the research works are SVM (Support Vector Machine) 40% (n=14/36), Decision Trees 40% (n=14/36), Regression techniques 36% (n=13/36), KNN (K-Nearest Neighbor) 22% (n=8/36) and Neural Network 20% (n=7/36).

The models are also analyzed based on evaluation metrics applied to the models for examining their performance.

The diagram in Fig. 5, depicts the percentage contribution of different evaluation metrics to ADR detection & prediction models. The precision & recall evaluation metric contributes to about 48% of ADR detection research papers while 60% of ADR prediction research works. The specificity, sensitivity & AUC are applied in about 30% of ADR detection research studies while only AUC is applied in about 70% of ADR prediction research papers. Lastly, the accuracy metric is used in around 40% of the research studies while the Ranking of drug-ADR association based on different metrics is involved in about 20% of the research works.

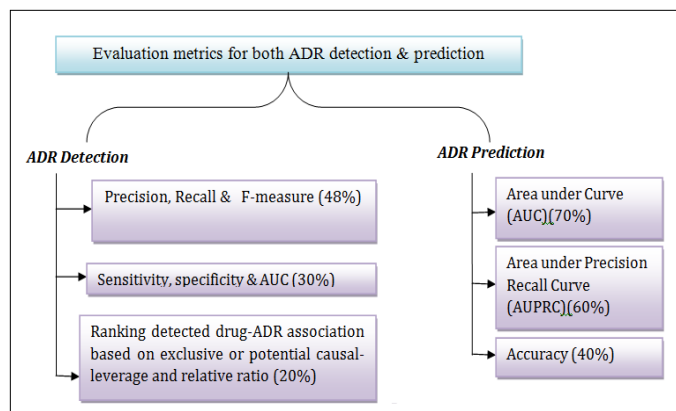


Fig. 5. Evaluation Metrics for ADR Detection and Prediction.

### C. Research Gaps Analysis

After reviewing the research studies, the author has identified some major dataset and technique-related limitations that are shown in the diagram:-

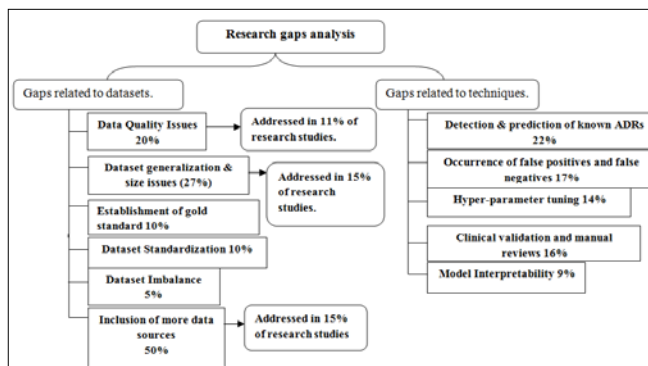


Fig. 6. Research Gap Analysis.

The Fig. 6, shows a thorough gap analysis depending on selected research papers. The significant research gaps related to ADR datasets are data quality, data generalization, and integration of more data sources which is specified in about 50% of the research papers. The research gaps are also analyzed based on the techniques applied for detecting and predicting an ADR. The major gaps discussed are the detection & prediction of known ADRs, the occurrence of false positives and false negatives, hyper-parameter tuning, and clinical validation. We have tried to address some research gaps in our research work but still many needs to be addressed for the future research study. These limitations form the basis to design our model for ADR prediction.

A framework is developed based on the gap analysis illustrated in Fig. 7.

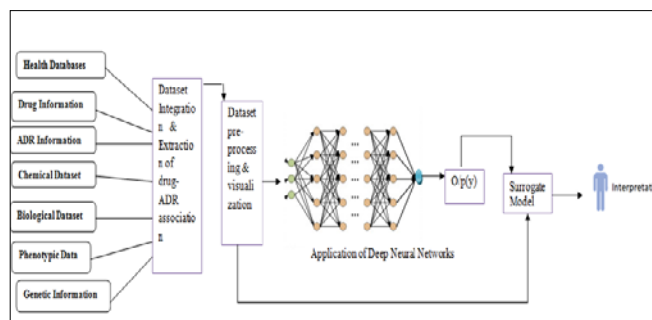


Fig. 7. Framework of the Proposed Model.

The framework signifies the key steps performed for ADR prediction. It tries to address the above-stated research gaps namely inclusion of more data sources, dataset imbalance, dataset size issues, and detection & prediction of known ADRs. The steps shown in the framework are practically implemented and results are derived accordingly.

### IV. DATASET SELECTION AND APPROACHES FOR INTEGRATION

The FAERS [23] data source used as input is a primary data source. It is available and freely accessible online. The data is collected and stored through an authentic process and

validated. This dataset is presented both in ASCII and CSV format. Around three million records were collected from the FAERS dataset dated from 2019 to 2020 end in ASCII format. Once downloaded and extracted the overall dataset is visualized in Fig. 8.

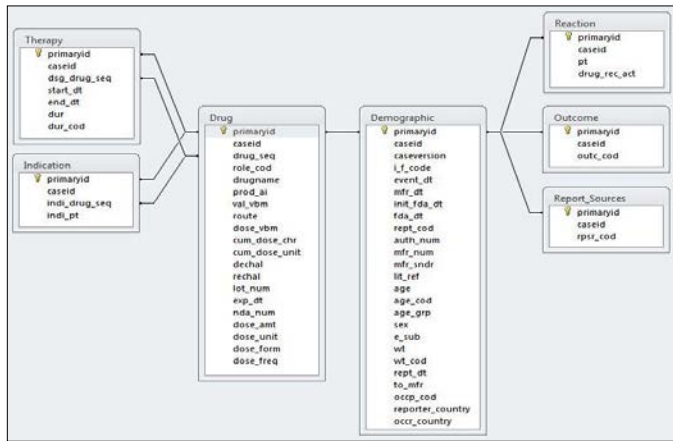


Fig. 8. FAERS Dataset [23].

The data in the FAERS dataset is unreadable and segregated across multiple tables. Therefore it is required to convert the dataset into a readable format and then integrate it using primaryid and caseid. The size of the integrated dataset is huge therefore it is necessary to detect and identify significant drug-ADR associations. The disproportionality analysis technique is applied for extracting such associations from the dataset.

The drug-ADR association is calculated in terms of PRR (Proportionality Reporting Ratio) [37]. Only those associations which are greater than the threshold value i.e.  $PRR \geq 3$  are filtered for further processing.

The output of the ADR detection algorithm is illustrated in Table II:-

TABLE II. RESULTS OF ADR DETECTION

Product	Adverse Event	Count	p_value	PRR
cc-10004	lymph node tuberculosis	102	-6167.969382	1.04847E+13
rifampicin	skin papilloma	6	-6168.662469	5.24282E+12
rapamune	hemiplegia	8	-6168.662467	5.24279E+12
alpelisib	osteonecrosis of jaw	16	-6168.662472	5.24272E+12

Overall the result of the initial experiments provides us with a processed and filtered FAERS dataset which is used further for integration with other data sources. The final ADR prediction is performed using drug characteristics as well as patient characteristics.

SIDER contains information regarding the marketed medicines along with their recorded ADRs. This dataset is secondary and is easily available on the internet for research purposes. The data source also includes information about drug indications on patients which are extracted from

unstructured text using NLP tools and techniques [27]. These drug indications help to distinguish ADRs from symptoms of disease and thus reducing the number of false positives. It is one of the most popular datasets used in ADR detection & prediction-based research study. It has been used in almost 60% of the research work done.

DrugBank was created by the University of Alberta and The Metabolomics Innovation Centre in Alberta, Canada [30]. It is a comprehensive, easily available, online data source that includes data about drugs and the protein targets of drugs. It also includes the components of proteins in terms of enzymes, transporters, receptors, and ion channels. The biological effect of drugs in terms of drug toxicity is also included as part of this research dataset. This dataset was acquired after obtaining the required permission from the authorities and assuring them of its ethical use.

PubChem includes information about drug molecules along with their chemical composition and their effect in response to the biology of patients. This data source is developed by NCBI (National Center for Biotechnology Information) which is a part of NLM (National Library of Medicine). The NLM is also included as a part of NIH (National Institutes of Health) of the USA [31]. It is also easily available for research purposes online.

The selected datasets are integrated using two different techniques. Each technique is illustrated in the following Fig. 9 and 10.

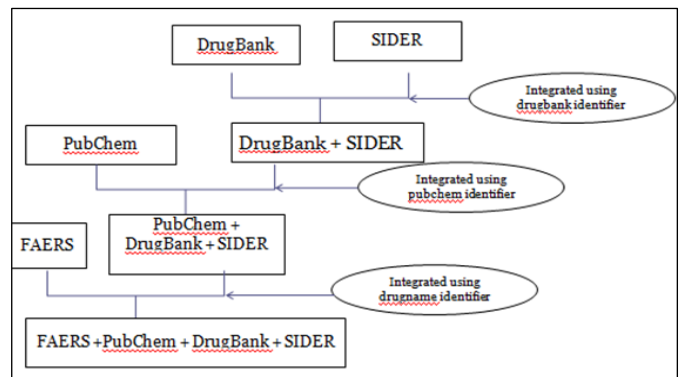


Fig. 9. Drug Identifier-based Integration.

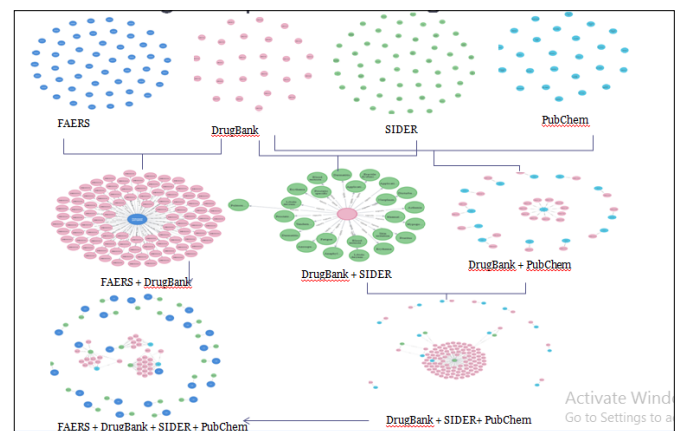


Fig. 10. Knowledge Graph Integration [38].



In drug identifier-based integration the DrugBank and SIDER are integrated using the drugbank identifier. Further PubChem is integrated using the PubChem identifier and lastly, FAERS is included using drug names. Similarly, in the case of knowledge graph [38] based integration, knowledge graphs are constructed using nodes of different datasets like drugs, target proteins, enzymes, pathways, indications, and adverse drug reactions. In the above figure, knowledge-graph integration information is derived from knowledge graphs which are used for identifying side-effects as well as detecting probable ADR for the prescribed medicines.

Further, the features of the integrated datasets are reviewed by the domain expert, and useful feedback and inputs were obtained by the author accordingly. Some features were dropped from the dataset while some were added as per their recommendations. The feature variables included as part of the integrated dataset are the type of target and target sequence. The type of targets can be divided into four categories receptors, ion channels, enzymes, and carrier molecules. Target sequences are genetic variants targeted by a given drug molecule. Apart from target type and target sequence some other features were also added as part of this dataset which is described in Table III.

TABLE III. FEATURE VARIABLE DESCRIPTION

Feature Variable	Description
LogP	Lipophilicity is a valuable parameter of the drug which affects its activity in the human body. The Log P value of the compound indicates the permeability of the drugs to reach the target tissue in the body[39]
LogS	The aqueous solubility of a compound significantly affects its absorption and distribution characteristics. Typically, a low solubility goes along with a bad absorption, and therefore the general aim is to avoid poorly soluble compounds. Our estimated logS value is a unit stripped logarithm (base 10) of the solubility measured in mol/liter.[39]
CYP inhibitors	The inhibitors are responsible for delaying the action of target proteins and that lead to a large amount of drug disposition in the human body which is harmful and severe.
Toxicity	The toxic nature of the drug molecule on the human body.

A. Dataset Preprocessing

The integrated dataset contains several redundant columns, null values, and categorical feature variables that must be pre-processed before further analysis. The steps involved in the pre-processing of both identifiers integrated and knowledge graph integrated datasets are described in the following Fig. 11.

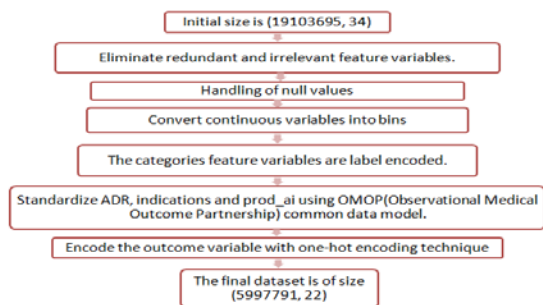
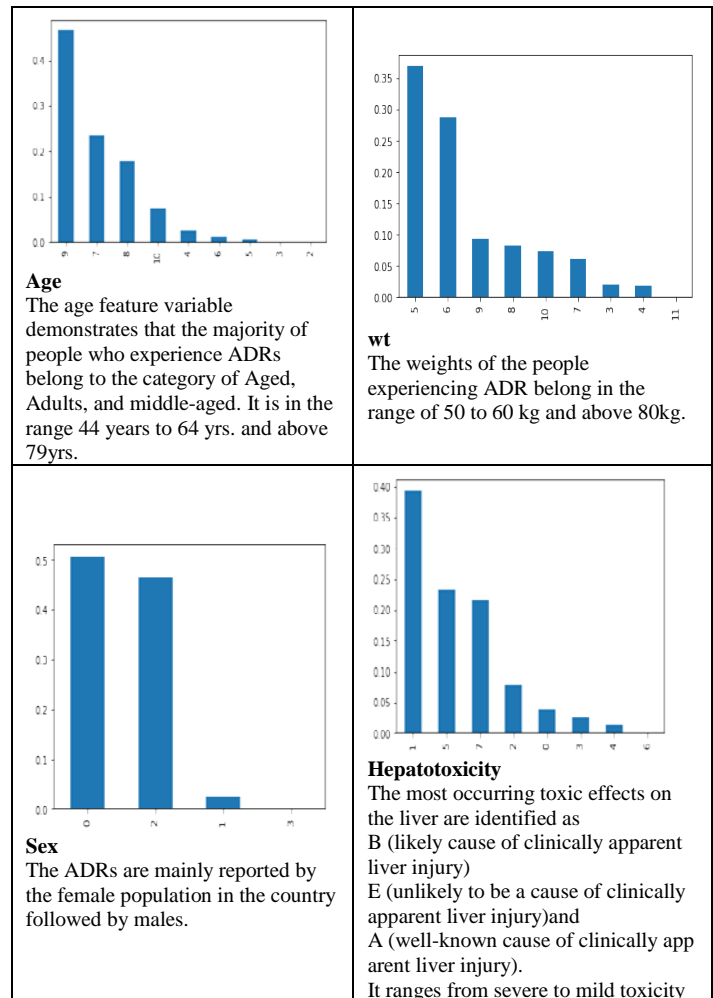


Fig. 11. Pre-processing Steps on Integrated Dataset.

The data distribution of the significant feature variables in the integrated dataset is visualized in the following bar charts shown in Table IV.

TABLE IV. DISTRIBUTION OF FEATURE VARIABLES



B. Kruskal Wallis Test

This test is used to determine whether or not there is a statistically significant difference between the medians of three or more independent groups [40]. The result of this test is shown in the output below.

```

H-statistic: 84253398.14222576
P-value: 0.0
Reject NULL hypothesis - Significant differences exist between groups.
    
```

The p-value is zero which is less than 0.05 which shows that a significant difference exists between groups and rejects the NULL hypothesis.

C. ADR Prediction Dataset

The current dataset includes only positive ADR samples. For any prediction problem, a balance of positive and negative data samples is required. Therefore the author applies GANs (Generative Adversarial Networks) [41] architectures to the original dataset and generates negative data samples based on the features of the original dataset. The output of the application of GANs is shown in Fig. 12.

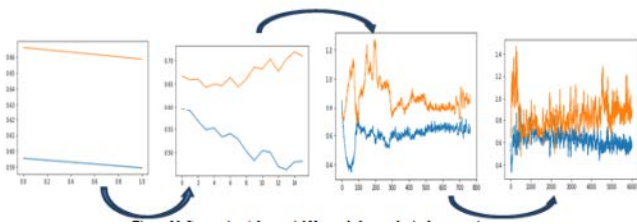


Fig. 12. Generative Adversarial Network for Data Creation.

A combined dataset of 20 lakh records was generated for both presence and absence of ADRs. This dataset will be used for the implementation of ADR prediction algorithms. The target class distribution before and after the application of GANs is illustrated in Fig. 13.

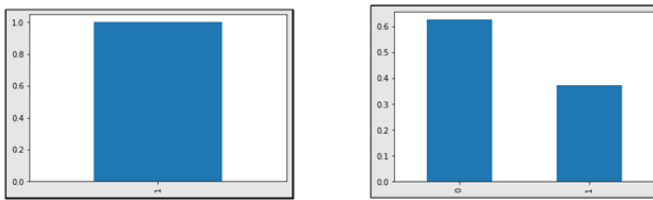


Fig. 13. Target Class Distribution.

The target class distribution shows the balance of positive and negative samples in the ADR dataset after the application of the GAN model.

## V. RESULTS

Deep learning models are Deep Neural Networks (DNN) containing non-linear processing units that transform raw data into higher-level representative information [40]. In recent years these techniques are actively applied in the field of drug discovery, precision medicine, protein engineering, genetic expression data analysis, and pharmacodynamics modeling [42]. Given the significant contribution of deep learning techniques in the domain of drug discovery, its capability can be very well extended to predict adverse reactions to drugs in humans. As previously discussed the ADR data sources both structured and unstructured are huge, diverse, and heterogeneous. DNN can successfully be applied to these data sources without the need for manual tuning. The initial training using a deep neural network is very complex and time-consuming but the network improves its performance by learning from input data.

Therefore the author proposes to apply deep learning models to the integrated dataset and evaluate its performance in terms of different evaluation metrics like accuracy, precision, recall, and F1 score. The model training is performed based on drug-ADR associations and other associated information. Deep learning performs well on a huge dataset. It also eliminates the need for hyper-parameter tuning. The number of hidden layers is optimized to provide the best results on the given dataset. The results obtained are shown in Table V:-

TABLE V. DEEP NEURAL NETWORK RESULTS

	Accuracy	Precision	Recall	F1 Score
The model with one hidden layer	0.57	0.9	0.57	0.64
The model with two hidden layer	0.91	0.92	0.91	0.91
The model with three hidden layer	0.55	1	0.55	0.71

From the results obtained it can be observed that the performance of the model with two hidden layers provides is optimum for all evaluation metrics. The performance seems to be consistent for precision but it varies significantly for the other evaluation metrics. It gives poor performance for one and three hidden layers but the best performance for two hidden layers. The results can be visualized in below Fig. 14.

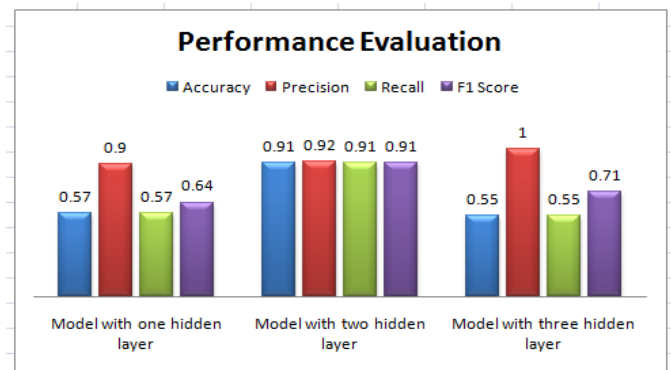


Fig. 14. Results Obtained by Models based on different Numbers of Hidden Layers.

Fig. 14 reflects the results obtained in the table and the performance of all evaluation metrics is consistent for the model with two hidden layers. It can be observed that the benefits of deep learning approaches are extensive but they suffer from the issue of non-interpretability. The 'black box' nature of deep learning techniques has restricted the interpretability of the model. The author has demonstrated the need for interpretable models for overall acceptability in the medical domain. Therefore to address this limitation the author has proposed the application of LIME (Local interpretable model-agnostic explanations) [43] for model explainability.

LIME is a technique that approximates any black-box learning model with a local, interpretable model to explain each prediction. From the definition it can be understood that LIME provides approximate explanations to individual prediction instances i.e. it is a local surrogate model. But to interpret the results based on the entire dataset SP-LIME [43] is applied. SP-LIME (Sub-modular Pick- Local Interpretable Model-Agnostic Explanation) tries to provide an answer to the question of developing trust for a given model for its acceptance. The trust is developed by dividing a given problem into several sub-problems for optimization. That means it identifies a series of instances along with their predictions that reflects the overall performance of the model based on the given data. The instances are selected in such a

manner that the features which are responsible for explaining different predictions are given higher importance value.

The results obtained by applying the SP-LIME algorithm to the given dataset are shown as follows:-

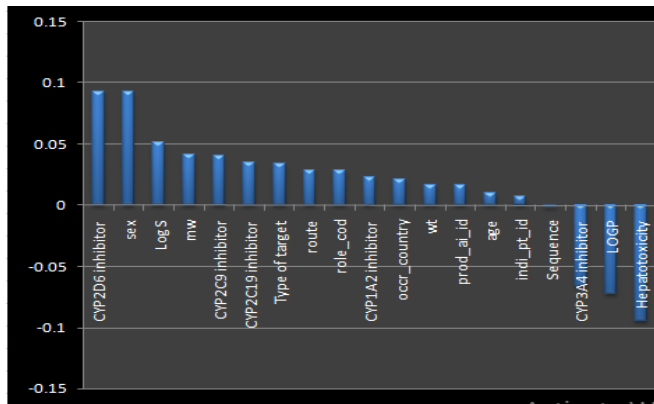


Fig. 15. SP-LIME Results.

Fig. 15 describes that the feature variable CYP2D6 inhibitor and sex contribute the highest to the target outcome prediction positively while hepatotoxicity, LOGP, and CYP3A4 are responsible for predicting the ADR outcome negatively. The proposed framework tries to address the research gaps stated in the existing research works. The inclusion of more data sources is identified in about 50% of the research studies, this issue is tackled by our proposed model. Other than this the model is trained on only validated drug-ADR association which is stated in about 16% of the research work. Lastly, the model's interpretability issue is also handled using a surrogate model. Therefore the proposed framework tries to address issues related to both data sources and techniques applied to these data sources.

## VI. DISCUSSION

Many reviews and surveys have been done in the past to address the issue of drug safety and healthcare. In 2015, Lardon et al. [11] in their research study tried to explore the breadth of evidence about the use of social media as a new source of knowledge for pharmacovigilance. They adopted a similar methodology of collecting research articles based on research questions and then analyzing them from multiple perspectives. The scope of their work is satisfactory but they have limited themselves to only unstructured datasources and NLP (Natural Language Processing) tools and techniques while in our research study the author has provided a comprehensive approach in terms of dataset selection and tools and techniques applied to them. Another research study was done by Ho et al. [44] in 2016 collected and analyzed research papers in terms of their problem statement, the dataset used and the methodology applied. The research summary provided in this paper is sufficient but it does not lead to any concrete solution to the existing research problem. Similarly, research studies conducted by Tan et al. [45] in 2016 have reviewed the interaction of different ADR datasets with biological and genetic datasets. Further, they discussed the benefits and limitations of these integrated datasets in the current scope. The drug-ADR associations are analyzed

statistically only on the basis datasource but no practical implementation is provided unlike in our research study. Although many other reviews and survey reports have discussed the major datasources related to ADR and their transition from a data-driven approach to machine learning models [42] for ADR prediction they do not provide an overall broad approach in terms of the datasources discussed, methodologies applied and a practical solution to the problem in the existing research works. Thus, our research study not only provides a comprehensive framework for both datasources and techniques applied to them but also implements the proposed model to obtain better results in terms of accuracy, F1 score, and interpretability.

## VII. CONCLUSION

In conclusion, this research study provides a bird's eye view of drugs, the importance of drug-ADR association, and the methodologies used to discover them. It also analyses its impact on human health. Although each step in this research study has been carried out in detail starting from research paper selection to proposed framework implementation and results in discussion, still some research gaps in the given study that should be considered for future research. First, the research papers are selected based on single drug-ADR association while research studies considering drug-drug interactions are ignored, so for future research work research studies considering drug-drug interaction should also be included. The proposed model has applied only a deep neural network for prediction and evaluated its performance based on the different number of hidden layers. Further, the author proposes to apply different deep learning models to this integrated dataset and then compare its performance with the existing results. The main aim of this research is to optimize the performance of the proposed model in terms of accuracy and other evaluation metrics.

## ACKNOWLEDGMENT

We would like to extend our sincere gratitude to Prof. Shreerang V. Joshi and Prof. Prashant S. Kharkar of the Institute of Chemical Technology (ICT), Mumbai as domain experts for their invaluable assistance and support in helping us to understand the pharmacological aspect of our research work. Their insights into our research have greatly improved the manuscript.

## REFERENCES

- [1] DL. Patrick, et al., "Patient-reported outcomes to support medical product labeling claims: FDA perspective." Value in health : the journal of the International Society for Pharmacoeconomics and Outcomes Research, vol. 10, Suppl 2, pp. S125-37, 2007.
- [2] L. Hazell, Lorna, and SAW. Shakir, "Under-reporting of adverse drug reactions : a systematic review." Drug safety vol. 29,no. 5, pp. 385-96, 2006.
- [3] The ICH Expert Working Group. (Nov. 2003). Post-approval safety data management: Definitions and standards for expedited reporting.
- [4] J.Lazarou, et al., "Incidence of adverse drug reactions in hospitalized patients: a meta-analysis of prospective studies." JAMA vol. 279,no.15, pp.1200-5,1998.
- [5] DK.Wysowski and L. Swartz, "Adverse Drug Event Surveillance and Drug Withdrawals in the United States, 1969-2002: The Importance of Reporting Suspected Reactions". Archives of Internal Medicine. Vol.165,no.12,pp.1363-9, 2005.

- [6] "Sibutramine (brand name Reductil) Information – Australia". Abbott Laboratories. 2010. Archived from the original on 2010-10-14. Retrieved 2010-10-08.
- [7] J.Jin, et al., "Factors affecting therapeutic compliance: A review from the patient's perspective." *Therapeutics and clinical risk management* vol. 4,no.1, pp. 269-86,2008.
- [8] M. Alomar., "Factors affecting the development of adverse drug reactions". *Saudi Pharmaceutical Journal*, Volume 22, Issue 2, Pages 83-94,2014.
- [9] G.Simper, et al., "Physiology and Pathology of Drug Hypersensitivity: Role of Human Leukocyte Antigens". *Physiology and Pathology of Immunology*, edited by Nima Rezaei, IntechOpen, 2017.
- [10] H.Arksey and L. O'Malley, "Scoping studies: towards a methodological framework." *International journal of social research methodology* vol. 8, no. 1, pp. 19-32, 2005.
- [11] J.Lardon, R. Abdellaoui, et al., "Adverse drug reaction identification and extraction in social media: a scoping review." *Journal of medical Internet research* vol. 17, no. 7, pp. e4304,2015.
- [12] R.Liu and P. Zhang, "Towards early detection of adverse drug reactions: combining pre-clinical drug structures and post-market safety reports." *BMC medical informatics and decision making* vol. 19, no. 1, pp. 1-9,2019.
- [13] M.S.Islam, et al., "A Systematic Review on Healthcare Analytics: Application and Theoretical Perspective of Data Mining." *Healthcare (Basel, Switzerland)* vol. 6,no.2, pp. 54,2018.
- [14] R. Ietswaart, et al., "Machine learning guided association of adverse drug reactions with in vitro target-based pharmacology." *EBioMedicine* vol. 57, pp.102837, 2020.
- [15] NIH(National Library of Medicine), <https://www.nlm.nih.gov/bidline.html>.
- [16] PubMed, <https://pubmed.ncbi.nlm.nih.gov/about/>.
- [17] Office of Scholarly Communication (December 2016). "A social networking site is not an open access repository". University of California.
- [18] ResearchGate, <https://www.researchgate.net/topic/Publication/>.
- [19] MeSH (Medical Subject Headings), <https://www.ncbi.nlm.nih.gov/mesh/>.
- [20] J.C.Talbot, J C, and B S Nilsson, "Pharmacovigilance in the pharmaceutical industry." *British journal of clinical pharmacology* vol. 45,no.5, pp. 427-31,1998.
- [21] P. Biswas, Pipasha, "Pharmacovigilance in Asia." *Journal of Pharmacology and Pharmacotherapeutics* vol. 4, no. 1\_suppl, pp. S7-S19, 2013.
- [22] D.Pappa and LK. Stergioulas, "Harnessing social media data for pharmacovigilance: a review of current state of the art, challenges and future directions." *International Journal of Data Science and Analytics* vol.8, no.2, 113-135,2019.
- [23] JM. Banda, et al., "A curated and standardized adverse drug event resource to accelerate drug safety research." *Scientific data* vol.3, no. 1 pp. 1-11,2016.
- [24] Dal Pan G. J., Arlett P. R., "The US Food and Drug Administration-European Medicines Agency collaboration in pharmacovigilance: common objectives and common challenges", *Drug Saf.*, Vol. 38, pp. 13-15.
- [25] B.Hugman, "From the Uppsala monitoring centre." *Drug safety* 28, no. 7, pp. 645-646, 2005.
- [26] K.Häyrynen, K. Saranto, and P. Nykänen, "Definition, structure, content, use and impacts of electronic health records: a review of the research literature." *International journal of medical informatics* vol. 77, no. 5,333 pp. 291-304,2008.
- [27] M.Kuhn, et al., "The SIDER database of drugs and side effects." *Nucleic acids research* vol. 44, no.D1, pp. D1075-9,2016.
- [28] NP. Tatonetti, et al., "Data-driven prediction of drug effects and interactions." *Science translational medicine* vol. 4, no. 125, pp. 125ra31-125ra31, 2012.
- [29] MC Cai,Q. Xu, et al., "ADReCS: an ontology database for aiding standardization and hierarchical classification of adverse drug reaction terms." *Nucleic acids research* 43, no. D1, pp.D907-D913,2015.3.
- [30] DS. Wishart,et al., "DrugBank: a knowledgebase for drugs, drug actions and drug targets." *Nucleic acids research* vol.36, no. suppl\_1, pp. D901-D906,2008.
- [31] Y. Wang, et al., "PubChem: a public information system for analyzing bioactivities of small molecules." *Nucleic acids research* vol. 37,S no. suppl\_2, W623-W633,2009.
- [32] VB Siramshetty, OA Eckert, et al., "SuperDRUG2: a one stop resource for approved/ marketed drugs." *Nucleic acids research* vol.346, no. D1,pp.D1137-D1143,2018.
- [33] M. Kuhn, et al., "STITCH: interaction networks of chemicals and proteins." *Nucleic acids research* vol 36, no. suppl\_1, pp.D684-D6883, 2007.
- [34] CF. Thorn, TE. Klein, and RB. Altman, "PharmGKB: the pharmacogenomics knowledge base." In *Pharmacogenomics*, pp. 311-320. Humana Press, Totowa, NJ, 2013.
- [35] M.Kanehisa and S. Goto,"Comprehensive gene and pathway analysis of cervical cancer progression." *Nucleic Acids Res* vol.28, pp. 327-30,2000.
- [36] DP. Hill, et al., "Gene Ontology annotations: what they mean and where they come from." In *BMC bioinformatics*, vol. 9, no. 5, pp. 1-9. BioMed Central, 2008.
- [37] A. Czarnecki and S. Voss, "Safety signals using proportional reporting ratios from company and regulatory authority databases." *Drug Information Journal* 42, no. 3, pp. 205-210,2008.
- [38] M.Wang,X. Ma,et al. "Adverse drug reaction discovery using a tumor-biomarker knowledge graph." *Frontiers in genetics* vol.11, pp. 625659, 2021.
- [39] NEH. Daoud, et al., "ADMET Profiling in Drug Discovery and Development: Perspectives of In Silico, In Vitro and Integrated Approaches." *Current Drug Metabolism* vol. 22, no. 7, pp. 503-522, 2021.
- [40] Y. Xia, "Correlation and association analyses in microbiome study integrating multiomics in health and disease." *Progress in Molecular Biology and Translational Science* vol. 171, pp. 309-491, 2020.
- [41] A. Creswell,et al., "Generative adversarial networks: An overview." *IEEE signal processing magazine* vol.35, no. 1, pp.53-65,2018.
- [42] CY.Lee and YP. Yen, "Machine learning on adverse drug reactions for pharmacovigilance." *Drug discovery today* vol. 24,no.7,pp. 1332-1343,2019.
- [43] R.ElShawi,Y. Sherif, M. Al-Mallah, and S. Sakr, "ILIME: local and global interpretable model-agnostic explainer of black-box decision." In *European Conference on Advances in Databases and Information Systems*, pp. 53-68. Springer, Cham, 2019.
- [44] TB. Ho,L. Le, DT. Thai, and S.Taewijit, "Data-driven approach to detect and predict adverse drug reactions." *Current pharmaceutical design* vol. 22, no. 23, pp. 3498-3526, 2016.
- [45] Y. Tan, Y. Hu, et al., "Improving drug safety: From adverse drug reaction knowledge discovery to clinical implementation." *Methods* vol. 110, pp.14-25, 2016.

# Secure Cloud Connected Indoor Hydroponic System via Multi-factor Authentication

Mohamad Khairul Hafizi Rahimi<sup>1</sup>

Department of Electrical, Electronic and Systems  
Engineering, Faculty of Engineering and Built Environment  
Universiti Kebangsaan Malaysia  
Bangi, Malaysia

Mohamad Hanif Md Saad<sup>2\*</sup>

Institute of IR 4.0 and Department of Mechanical and  
Manufacturing Engineering, Faculty of Engineering and  
Built Environment  
Universiti Kebangsaan Malaysia  
Bangi, Malaysia

Aini Hussain<sup>3</sup>

Department of Electrical, Electronic and Systems  
Engineering, Faculty of Engineering and Built Environment  
Universiti Kebangsaan Malaysia  
Bangi, Malaysia

Nurul Maisarah Hamdan<sup>4</sup>

Department of Mechanical and Manufacturing Engineering,  
Faculty of Engineering and Built Environment  
Universiti Kebangsaan Malaysia  
Bangi, Malaysia

**Abstract**—Now-a-days, the hydroponic farming system with the Internet of Things (IoT) technology is increasingly becoming a trend for researchers to produce a more capable farming device or remote monitoring system. However, this intelligent system is not controlled securely and will be dangerous when system hacking occurs. Therefore, developing a secure indoor hydroponic monitoring device with multi-factor authentication (MFA) method is proposed. The research aims to develop a secure cloud-connected indoor hydroponic system via multi-factor authentication on the ThingsSentral IoT platform with an MFA technique. The developed system comprises an iPhone Operating System (iOS), an Arduino node microcontroller unit and a ThingsSentral web IoT platform. A security software application on iOS phones with MFA techniques is built to authenticate devices before communicating with ThingsSentral.io. Token authentication between ThingsSentral.io and the security software application must be done before the hydroponic monitoring device can send and receive data. An indoor hydroponic monitoring system device with MFA security technique has been successfully produced from the study. An MFA security technique for iOS apps has also been successfully developed. In conclusion, using the MFA technique, this research successfully develops a high-security control and communication system between the field device and the IoT platform. Although the MFA security system developed for this IoT platform has several steps that need to be done before data can be sent to the cloud database, the users themselves can allow or prohibit a device from operating. Besides, users can also control and monitor the security of the device and the IoT platform when they operate.

**Keywords**—Internet of things; intelligent system; remote monitoring; hydroponic; multi-factor authentication

## I. INTRODUCTION

Due to an increase in population and a country's development, a diverse alternative has dealt with the crisis of

adequate food provision. This is because land and farms must be reduced to make room for more excellent homes. The development of indoor hydroponics farming, a relatively successful means of producing a crop, is one of the most recent agricultural technologies established to mitigate this problem [1]–[4]. Hydroponics technology is used throughout the world. Numerous methods and techniques have developed. Most hydroponic surrounding areas are prepared with their temperature, humidity, electrical conductivity (EC) and pH manually measured by producers. Adjustments are then calculated to meet the needs of the plants [5]–[7].

With the advent of IoT technology, hydroponic farming systems can increasingly thrive. The combination of IoT technology in this hydroponic farming system can also monitor and control the condition of the crop environment in real-time. In addition, it can also reduce inefficiencies and improve agricultural performance [8], [9]. Combining the Internet of Things and the agriculture industry can minimize inefficiencies in food production and expand the food market [10], [11]. The potential of IoT systems can connect breakthrough technology like plants that deploy sensors.

IoT technologies that use the medium of internet connection can send data to a cloud server or receive instructions to operate [12]. Most IoT technologies use an open internet connection medium to perform these processes. When an IoT device uses an open internet connection, it will expose to dangers such as information theft or system hacking [13].

Hydroponic farming systems combined with IoT technology are increasingly becoming a trend for researchers to produce a more capable farming device or system. However, studies related to secure hydroponic monitoring systems with IoT integration are still lacking. If this IoT hydroponic farming system is not controlled securely, it will be a danger when system hacking occurs. Hence to solve the problem, the current novelty research proposes the development of a secure indoor

---

This research is supporting from Ministry of Higher Education (MoHE) through the grant LRG5/1/2019/UKM-UKM/5/2 and Faculty of Engineering and Built Environment, Universiti Kebangsaan Malaysia (UKM)

\*Corresponding Author



hydroponic monitoring device with multi-factor authentication (MFA) method.

This study has developed a security authentication application on mobile phones to control the connection between the gateway device and the ThingsSentral application. Next, a gateway device is developed to send the data obtained from the sensors to the ThingsSentral application using the Hypertext Transfer Protocol (HTTP) communication protocol. The security authentication application only needs to be done at the beginning of the gateway device requesting to send data only.

## II. HYDROPONIC MONITORING SYSTEM

A literature review has been made related to the hydroponic monitoring system. This section describes some research done by other researchers related to electronic technology used in producing hydroponic monitoring systems. This section also shows the results obtained by past researchers during this study.

In the era of technology centuries, human lives have become more accessible in all parts with the growth of wireless technology, the Internet of Things (IoT). Agricultural systems' decisions demonstrate the rapid rise in internet users over the previous decade. The IoT development guarantees that the surveillance system technique becomes more advanced on the 'user's terms and can be accessed anytime and anywhere within a distant place [14]–[16]. The development of a hydroponic farming system communication based on the IoT platform monitors the plant condition at a remote monitoring station. It allows the automatic system to turn on when necessary [17]–[19].

Melchizedek I. Alipio et al. [20] employed a Bayesian network to analyze the data and automate the hydroponics method as the human population expands to meet the food requirement. Using IoT technology in conjunction with hydroponics allows a high-quality farmer to produce more. The farm's hydroponics is linked to the light intensity, pH, EC, water temperature, and humidity sensors. Sensor data were gathered for analysis, and a Bayesian Network was built. The nutrient film method (NFT) was used with sensors and actuators attached to the plant. To use a remote location to monitor and gather data from the hydroponics farm and deliver it to the web interface.

Fuzzy logic was utilized to regulate the supply of nutrients to hydroponic plants, according to M. Fuangthong and P. Pramokchon [21]. Instead of employing soil, water solvent was used to supply nutrients to the crops. Farmers in soilless culture need extra attention from growing plants in hydroponics to analyze nutrient solvent's EC and pH levels. The pH value varies depending on the crop, and the automatic monitoring mechanism provides a nutrient solution. A farmer's practical abilities are required to prevent an excess delivery of nutrient solutions to plants. The Dynamic Root Floating Technique (DRFT) automatically controls the nutrient solvent flow to hydroponic plants using fuzzy logic.

In 2017, Ms.S.Charumathi et al. [22] proposed a novel cultivating crops in soilless culture. Hydroponics can enhance yields while taking up less space and producing high-quality

crops. Because of the less fertile soil, farmers utilize pesticides and fertilizers to produce crops that damage human health. In a closed location, the hydroponics arrangement can alleviate traditional framing issues. The Arduino Microcontroller was combined with the IoT concept to detect the situation around the plants automatically. The proposed method has a major flaw in that it requires a farmer's observation because the data is stored locally.

R.Rajkumar and R.Dharmaraj [23] proposed a hydroponics idea with the wireless sensor network. Planting in a contaminated climate is difficult in conventional farming. Improving the yield requires more fertile soil. It also necessitates a higher weed extraction cost and a large growing area. Then, seasonal food processing is only possible at that time. The hydroponics technique allows for year-round production of the crops. Instead of using artificial mineral nutrients, the author experimented with hydroponics using organic ash fertilizer. Moisture, temperature, and water level are all sensed by the microcontroller. The sensor data is sent to the cloud, allowing the farmer to track the progress of his plants from afar. The author uses Blynk, an open-source API that receives sensor data.

M.K.R.Effendi et al. [24] developed smart farms and agriculture. The Internet of Things (IoT) and data analytics are being used to improve farm and agriculture sector operational efficiency and productivity. They devised monitoring, control, or automation system to help the farmer and then gathered all data on rainfall, temperature, humidity, and light intensity. The development included hardware, software, programming, and sensors such as a water sensor, light-dependent resistor sensor, temperature and humidity sensor, and weight sensor for data collection. The project's outcome is that they were able to apply IoT concepts to aquaponics and goat stall monitoring, control, and automation systems.

With the Wireless Sensor Network, Jumras Pitakphongmetha et al. [25] suggested a hydroponics technique to transfer sensor data to the cloud. Compared to conventional farming, hydroponics lets farmers raise money with higher yields. With global warming, the natural climate is challenging to foresee. Hydroponics can solve this issue by introducing without disturbance from the atmosphere in a protected environment. The significant parameter was then calculated using a different sensor around the hydroponics plant.

From the literature review, we found that the study related to hydroponic agriculture monitoring systems is sustainable and somewhat advanced in terms of controlling the environmental conditions of crops. However, most hydroponic monitoring devices are lacking from the perspective related to the security and safety of the monitoring devices. Electronic devices are vulnerable to a cyber-security threat and can be misused by others to cause harm to the community or society. These security and safety aspects are very much lacking and should be given extra attention. Although numerous security techniques and devices are available, most IoT devices do not come with security-enabled capabilities. Amongst the available and often used security methods for the connected or IoT system is via the authentication approach [26] and [27].



Security technologies can be implemented by integrating electronic devices, computers, and wireless communication to provide further protection. The second-factor authentication strategy, often utilized in account defense, is the next degree of protection [28], [29]. Two-factor authentication is a type of authentication that determines its identity using two out of three variables. “Something the user knows,” “something the user has,” and “something in the user” are three regularly seen variables [2], [30], [31]. The use of this two-factor authentication approach improves network security. This is due to the use of technological devices with an internet connection.

This combination of IoT and urban agriculture will evaluate data automatically by uploading it to the cloud and allowing users to make decisions [32], [33]. Furthermore, this indoor farming system is self-contained and can work safely with this network security mechanism. The ability to produce could change the agriculture industry, helping to enhance the smart but inefficient rural sector in our economy.

### III. SECURE INDOOR HYDROPONIC MONITORING SYSTEM DEVELOPMENT

This section describes in detail the methodology implemented in this research. The study of how people utilize hydroponic systems was carried out to create a monitoring device for the farming system. Furthermore, the proposed approach makes the hydroponic system simple to maintain. Multiple security can be achieved by using the general design of the system with an MFA approach. This section has two parts: indoor hydroponic device monitoring system design and device security system design.

#### A. Device Monitoring System Design

Fig. 1 shows the main layout of an indoor vertical hydroponic farming system using the MFA method. Indoor hydroponic devices include three major components: input, controller, and output. On the input side, water electrical conductivity (EC), water parts per million (PPM), humidity sensors, temperature sensors, and keypad (users to enter their Wi-Fi information) were used. Using a Node microcontroller in the controller portion is critical for securing the input part to the output and the input to the cloud database. Because NodeMCU acts as a WI-FI chipset, it can send data to the internet. An OLED panel displays data straight from the sensor on the output side. Fig. 2 shows the sensor and module information connected to the NodeMCU ESP8266. While Fig. 3 shows the indoor hydroponic monitoring device that has been connected.

The system starts with the user configuring the service set identifier (SSID) and the password to connect to the internet using the keypad provided. Once the SSID and password of the nearest internet have been successfully entered, the user needs to enter the user and device id, which can be obtained on the ThingsSentral website. Once the user and device id have been entered, a notification will be sent to the security verification application on the user’s phone. Fig. 4 below shows the interface of the security authentication application.

Users can choose either to allow or not the hydroponic monitoring device to function and be able to send data to the ThingsSentral platform. The National University of Malaysia

developed ThingsSentral, a cloud based IoT platform (UKM). The ThingsSentral application’s UI is shown in Fig. 5. This platform allows users to develop their own cloud based IoT systems for data collection, storage, and retrieval. This system platform communicates over the internet using the HTTP protocol. Data can be sent to or from hardware microcontrollers like Arduino, Node MCU, and Raspberry Pi. ThingsSentral’s functioning is based on channels that comprise data fields, position fields, and status fields.

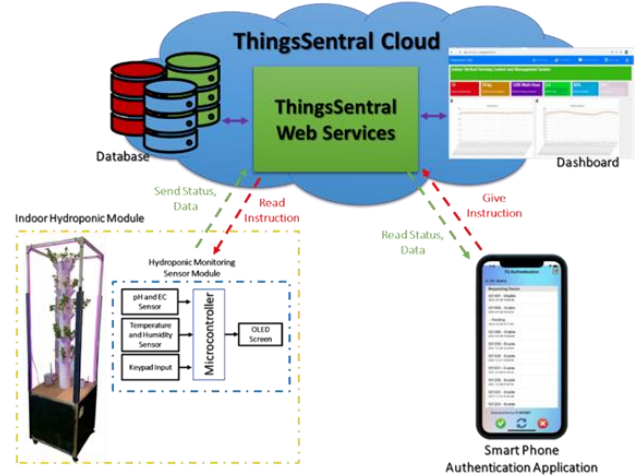


Fig. 1. System Design Framework of Hydroponics Farming System with an MFA Method.

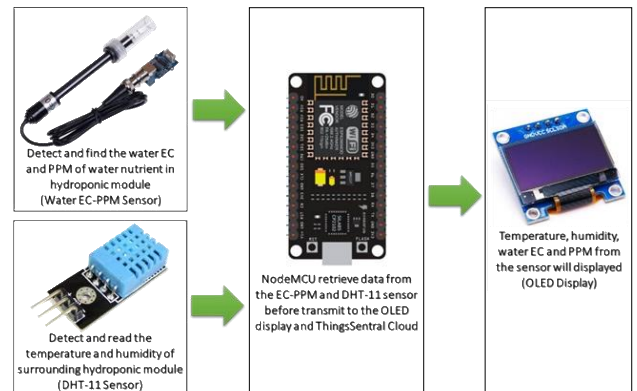


Fig. 2. The Required Hardware for Monitoring Temperature, Humidity, EC and PPM with Component Description [34]–[37].

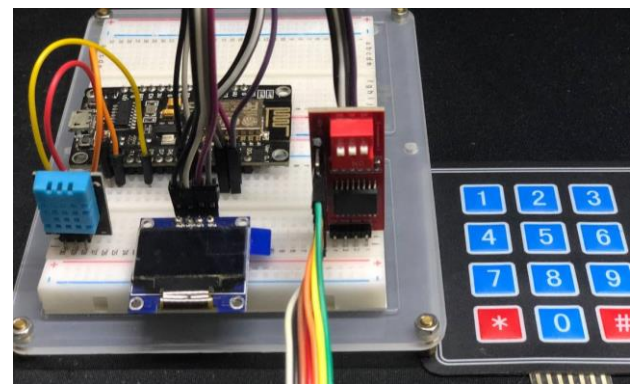


Fig. 3. The Hardware Set Up for the Indoor Hydroponic Monitoring System used in this Study.

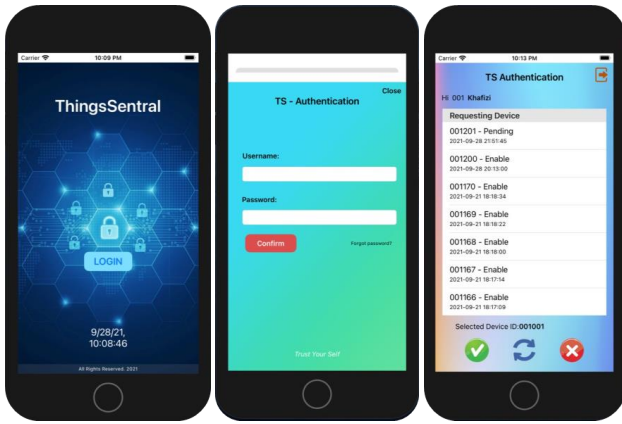


Fig. 4. An MFA Security Interface for Mobile Application.

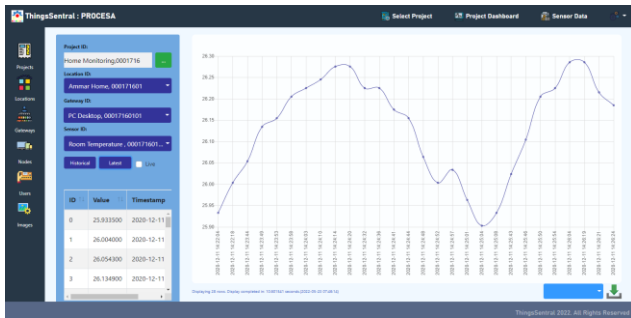


Fig. 5. ThingsSentral Application for Live Data Monitoring.

After the user confirms the indoor hydroponic monitoring device, it will read the temperature and humidity of the plant environment. The plant environment data can also be sent to the cloud database. Fig. 7 shows that the hydroponic model was built. While Fig. 6 shows the readings obtained by the ambient temperature and humidity monitoring device.

### B. Device Security System Design

This section describes the proposed multi-factor authentication that uses a credential token key. Fig. 8 shows the architecture of the proposed system. The system consists of an IoT device as a client, a ThingsSentral server as a web services platform, and an authentication application as a phone authentication application. All these systems need to perform registration with the ThingsSentral Server. ThingsSentral server generates unique registration IDs for all entities and certificates of credential token keys for IoT devices. In this system, the mutual authentication between IoT device and authentication services, authentication application and authentication services, token generator services, and IoT devices occur. The certificate of a credential token key

establishes between IoT devices and ThingsSentral primary services. The proposed security architecture system consists of five processes, which explain as follows.



Fig. 6. Indoor Vertical Hydroponic Model for a Data Monitoring System.

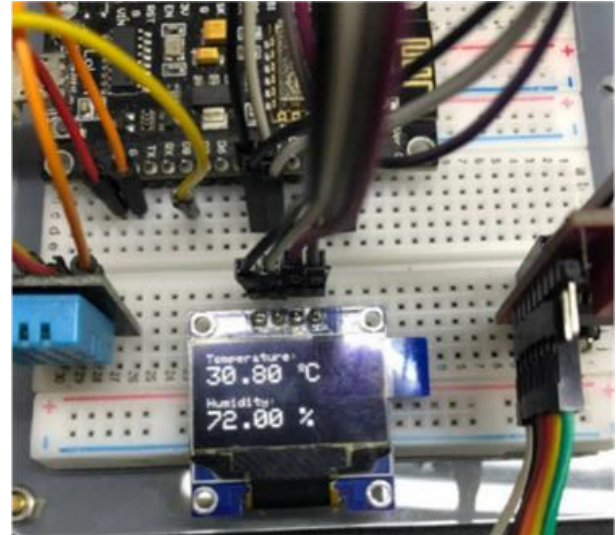


Fig. 7. Live Data Pickup by Sensor Shown in I2C OLED Display.

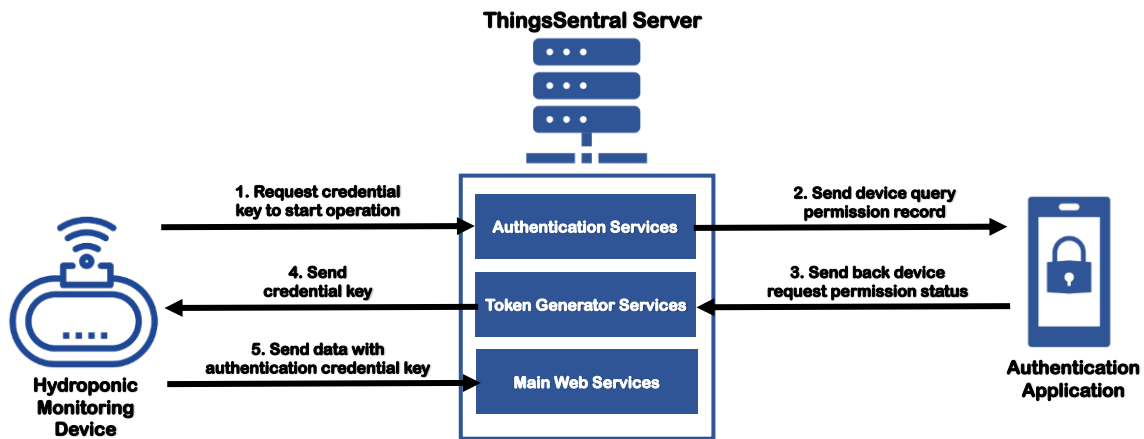


Fig. 8. The Proposed Architecture for the Security System.

IV. RESULTS AND DISCUSSION

The results obtained after applying the methods are described presently. The results relate to the indoor hydroponic farming monitoring module and hardware security system using the MFA authentication method.

In order to obtain this result, the study's implementation method and prototype model have been shown in Fig. 2, 3 and 7. Fig. 9 represents the screenshot of the authentication process performed from the start of the monitoring device until the data was retrieved from the water nutrient and DHT-11 sensor. In contrast, Fig. 10 and Fig. 11 show the screenshots of the dashboard display for the IoT-based hydroponic crop monitoring system on a PC/laptop and a mobile phone, respectively. As can be seen, the ThingsSentral platform has afforded an informative, real-time and excellent visualization approach to monitoring the hydroponic crop. The IoT-based

monitoring system has successfully implemented a secured system using multi-factor authentication.

The local OLED display module and the centralized cloud dashboard on ThingsSentral will display all data obtained at the hardware module. Then it is displayed on the I2C OLED. The I2C OLED provided the interface for user WI-FI SSID, WI-FI password, user id and device gateway. The water nutrient solution (EC and PPM), humidity and temperature of the surrounding hydroponic module were also displayed on the OLED. The data analysis was based on the data presented in the methodology and also the data obtained from the graph on ThingsSentral, as shown in Fig. 10 and Fig. 11. The real-time timestamps and data for the indoor hydroponic monitoring module on the water nutrient solution and surrounding humidity and temperature were recorded on the ThingsSentral platform. The result for the MFA validation application display is shown in Fig. 4.

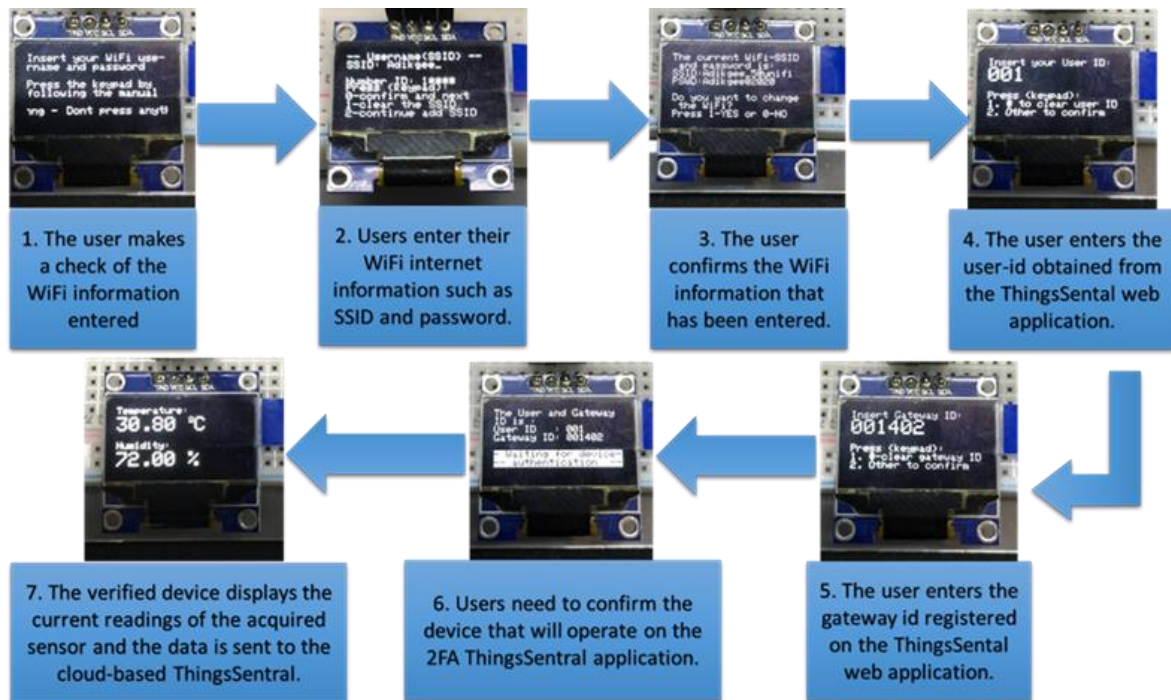


Fig. 9. Dashboard for ThingsSentral Web-based Interface with MFA Method.



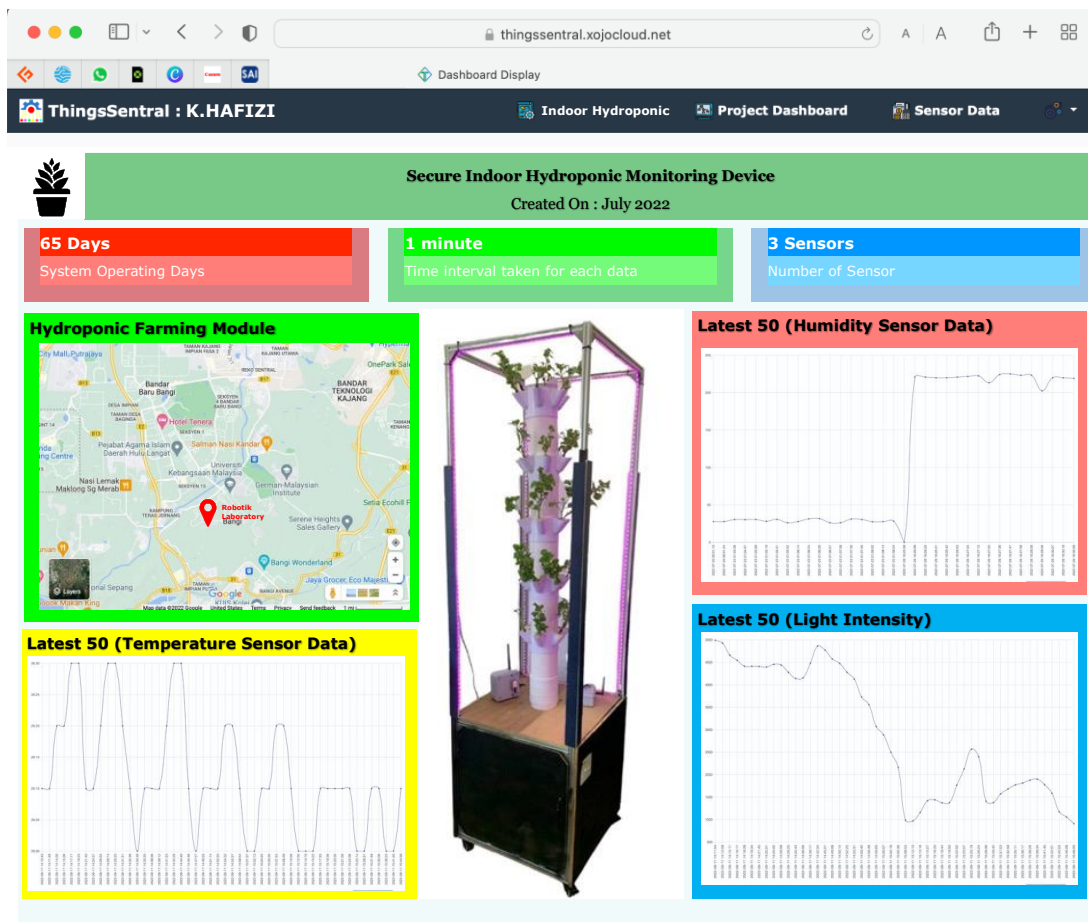


Fig. 10. The OLED Displays Wi-Fi and Device Monitoring Registration Process Flow.



Fig. 11. MFA Method used on a Dashboard for ThingsSentral Web-based Interface Viewed on Mobile Phones.

The hydroponic monitoring module was tested 100 times using proposed IoT platforms. Each time this monitoring device successfully displays the indoor hydroponic module system’s current data, this device will be turned off and on again to repeat the device verification process and data transmission to the cloud database. Fig. 12 shows the time graph for 100 times hydroponic monitoring devices to obtain the credential key and send data to the ThingsSentral server. From the data obtained, the average time taken for this system to start operating until the data received on the ThingsSentral server is 579ms (red line). The results for the average time required have been included in Table I and compared with other platforms.

Table I shows the names of the different IoT platforms, the average time taken and their respective security techniques. As result of the comparison that has been made shows that the time taken by the ThingsSentral IoT platform with the MFA technique takes a little longer compared to other IoT platforms. However, the ThingsSentral IoT platform uses the MFA technique that allows users to authenticate a device to operate. ThingsSentral also uses a dynamic credential key every time the device starts operating. Unlike other IoT platforms, they use API keys or token keys only. In addition, this other platform uses a static token key. If hackers or strangers can figure out this token key, they can do unexpected things.

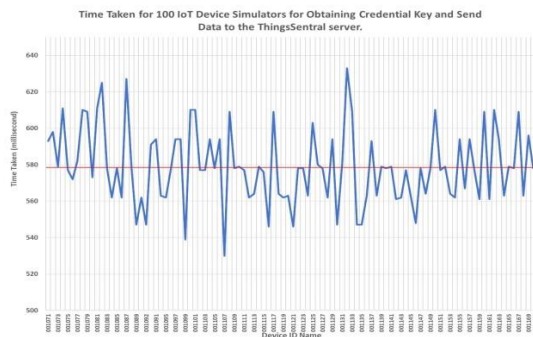


Fig. 12. Time Graph for 100 Times IoT monitoring Device Tested Request Credential Key and Send Data to ThingsSentral Server.

TABLE I. COMPARISON OF SYSTEM SECURITY FOR FIVE DIFFERENT IOT PLATFORMS

IoT Platform Name	Security Technique	Average Time Taken / Response Time (ms)
Kaa IoT [33]–[38], [40]–[42]	SSL security elements are combined with basic authentication through JSON web tokens.	150
ThingSpeak [43]–[47]	Secured MQTT broker and random static token on personal API Key	25.2
Thingier.io [39], [48]–[51]	SSL security elements are combined with basic authentication through JSON web tokens.	266.7
Thingsboard [39], [52]–[55]	Encryption algorithms on SSL and credential types certificates and access tokens.	217.5
ThingsSentral (proposed platform)	SSL security features on web API services with a multi-factor authentication method on the registered device.	579

## V. CONCLUSIONS

This paper describes developing a secure indoor hydroponic farming monitoring system based on a wireless system using the XOJO platform, ThingsSentral IoT, NodeMCU, and hydroponic sensors. The hydroponic monitoring system device was tested on an indoor vertical hydroponics module and the system’s functionality was assessed. The system can display temperature, humidity, and nutrient solution content in water. The developed system also uses the MFA method to increase further the level of communication between the monitoring device and the cloud database. This system is also compared with several other cloud database IoT platforms. One noteworthy finding is that this developed system takes a little longer than the use of cloud database IoT platforms. However, this system is safer because the ThingsSentral IoT platform uses authentication from humans or owners to the device itself. Unlike other IoT platforms, they only use the API Key or token that needs to be entered into the device. This study only used one indoor vertical hydroponic module to obtain data. Therefore, the data used is only to measure the security and usage of cloud based IoT platforms at a time. In the future, studies on several indoor vertical hydroponic modules can be used to obtain more data to produce more accurate and valuable results. The data obtained from the hydroponic monitoring module can be compared with the existing device. In addition, future research can be done by developing more hydroponic monitoring sensors to control plant environmental conditions and a phone application by integrating it with the data obtained from the device. The data obtained will be displayed on the user’s phone application. Users can also see their crops’ condition from time to time, even if they are far from the crops.

## ACKNOWLEDGMENT

The authors would like to thank the Ministry of Higher Education (MoHE) for supporting this research through the LRGS/1/2019/UKM-UKM/5/2 grant code and Faculty of Engineering and Built Environment, Universiti Kebangsaan Malaysia (UKM) for providing the research facilities.

## REFERENCES

- [1] N. Engler and M. Krarti, “Review of energy efficiency in controlled environment agriculture,” *Renew. Sustain. Energy Rev.*, vol. 141, p. 110786, 2021.
- [2] M. K. H. Rahimi, M. H. M. Saad, A. H. M. Juhari, M. K. A. M. Sulaiman, and A. Hussain, “A Secure Cloud Enabled Indoor Hydroponic System Via ThingsSentral IoT Platform,” in *2020 IEEE 8th Conference on Systems, Process and Control (ICSPC)*, 2020, pp. 214–219.
- [3] M. H. Tunio, J. Gao, S. A. Shaikh, I. A. Lakhari, W. A. Qureshi, K. A. Solangi, & F. A. Chandio, “Potato production in aeroponics: An emerging food growing system in sustainable agriculture for food security,” *Chil. J. Agric. Res.*, vol. 80, no. 1, pp. 118–132, 2020.
- [4] F. A. A. Khan, “A review on hydroponic greenhouse cultivation for sustainable agriculture,” *Int. J. Agric. Environ. Food Sci.*, vol. 2, no. 2, pp. 59–66, 2018.
- [5] A. Dutta, I. Nag, S. Basu, D. Seal, and R. K. Gayen, “IoT based Indoor Hydroponics System,” in *2021 5th International Conference on Electronics, Materials Engineering & Nano-Technology (IEMENTech)*, 2021, pp. 1–8.
- [6] S. Khan, A. Purohit, and N. Vadsaria, “Hydroponics: current and future state of the art in farming,” *J. Plant Nutr.*, vol. 44, no. 10, pp. 1515–1538, 2020.

- [7] M. S. Kamarulzaman, S. A. Jumaat, M. N. Ismail, and A. F. M. Nor, "Monitoring System of Hydroponic Using Solar Energy," *J. Electron. Volt. Appl.*, vol. 2, no. 1, pp. 26–37, 2021.
- [8] Z. Xu, A. Elomri, T. Al-Ansari, L. Kerbache, and T. El Mekki, "Decisions on design and planning of solar-assisted hydroponic farms under various subsidy schemes," *Renew. Sustain. Energy Rev.*, vol. 156, p. 111958, 2022.
- [9] S. Goddek and K. J. Keesman, "Improving nutrient and water use efficiencies in multi-loop aquaponics systems," *Aquac. Int.*, vol. 28, no. 6, pp. 2481–2490, 2020.
- [10] A. A. R. Madushanki, M. N. Halgamuge, W. A. H. S. Wirasagoda, and S. Ali, "Adoption of the Internet of Things (IoT) in agriculture and smart farming towards urban greening: A review," *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 4, 2019.
- [11] D. Mao, Z. Hao, F. Wang, and H. Li, "Innovative blockchain-based approach for sustainable and credible environment in food trade: A case study in shandong province, china," *Sustainability*, vol. 10, no. 9, p. 3149, 2018.
- [12] L. M. Abdulrahman, S. R. Zeebaree, S. F. Kak, M. A. Sadeeq, A. Z. Adel, B. W. Salim, & K. H. Sharif, "A state of art for smart gateways issues and modification," *Asian J. Res. Comput. Sci.*, pp. 1–13, 2021.
- [13] C. L. Stergiou, K. E. Psannis, and B. B. Gupta, "IoT-based big data secure management in the fog over a 6G wireless network," *IEEE Internet Things J.*, vol. 8, no. 7, pp. 5164–5171, 2020.
- [14] H. Tabrizchi and M. Kuchaki Rafsanjani, "A survey on security challenges in cloud computing: issues, threats, and solutions," *J. Supercomput.*, vol. 76, no. 12, pp. 9493–9532, 2020.
- [15] P. M. Chanal and M. S. Kakkasageri, "Security and privacy in IOT: a survey," *Wirel. Pers. Commun.*, vol. 115, no. 2, pp. 1667–1693, 2020.
- [16] S. Mishra and A. K. Tyagi, "The Role of Machine Learning Techniques in Internet of Things-Based Cloud Applications," in *Artificial Intelligence-based Internet of Things Systems*, Springer, 2022, pp. 105–135.
- [17] R. Rayhana, G. Xiao, and Z. Liu, "Internet of things empowered smart greenhouse farming," *IEEE J. Radio Freq. Identif.*, vol. 4, no. 3, pp. 195–211, 2020.
- [18] O. Friha, M. A. Ferrag, L. Shu, L. A. Maglaras, and X. Wang, "Internet of Things for the Future of Smart Agriculture: A Comprehensive Survey of Emerging Technologies," *IEEE CAA J. Autom. Sin.*, vol. 8, no. 4, pp. 718–752, 2021.
- [19] M. S. Farooq, S. Riaz, A. Abid, K. Abid, and M. A. Naem, "A Survey on the Role of IoT in Agriculture for the Implementation of Smart Farming," *IEEE Access*, vol. 7, pp. 156237–156271, 2019.
- [20] M. I. Alipio, A. E. M. Dela Cruz, J. D. A. Doria, and R. M. S. Fruto, "A smart hydroponics farming system using exact inference in Bayesian network," *2017 IEEE 6th Glob. Conf. Consum. Electron. GCCE 2017*, vol. 2017-Janua, no. Gccee, pp. 1–5, 2017, doi: 10.1109/GCCE.2017.8229470.
- [21] M. Fuangthong and P. Pramokchon, "Automatic control of electrical conductivity and PH using fuzzy logic for hydroponics system," *3rd Int. Conf. Digit. Arts, Media Technol. ICDAMT 2018*, pp. 65–70, 2018, doi: 10.1109/ICDAMT.2018.8376497.
- [22] S. Charumathi, R. M. Kaviya, J. Kumariyarsi, R. Manisha, and P. Dhivya, "Optimization and Control of Hydroponics Agriculture using IOT," *Asian J. Appl. Sci. Technol.*, vol. 1, no. 2, pp. 96–98, 2017, [Online]. Available: [https://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=2941105](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2941105).
- [23] R. Rajkumar, R. Dharmaraj, and P. Scholar, "A Novel Approach for Smart Hydroponic Farming Using IoT," *Int. J. Eng. Res. Comput. Sci. Eng.*, vol. 5, no. 5, pp. 18–23, 2018, [Online]. Available: [https://www.technoarete.org/common\\_abstract/pdf/IJERCSE/v5/i5/Ext\\_71306.pdf](https://www.technoarete.org/common_abstract/pdf/IJERCSE/v5/i5/Ext_71306.pdf).
- [24] M. K. R. Effendi, M. Kassim, N. A. Sulaiman, and S. Shahbudin, "IoT Smart Agriculture for Aquaponics and Maintaining Goat Stall System," *Int. J. Integr. Eng.*, vol. 12, no. 8, pp. 240–250, 2020.
- [25] J. Pitakphongmetha, N. Boonnam, S. Wongkoon, T. Horanont, D. Somkiadcharoen, and J. Prapakornpilai, "Internet of things for planting in smart farm hydroponics style," *20th Int. Comput. Sci. Eng. Conf. Smart Ubiquitous Comput. Knowledge, ICSEC 2016, 2017*, doi: 10.1109/ICSEC.2016.7859872.
- [26] L. Tawalbeh, F. Muheidat, M. Tawalbeh, and M. Quwaider, "IoT Privacy and security: Challenges and solutions," *Appl. Sci.*, vol. 10, no. 12, p. 4102, 2020.
- [27] D. Rahadiyan, S. Hartati, and A. P. Nugroho, "Design of an Intelligent Hydroponics System to Identify Macronutrient Deficiencies in Chili," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 1, 2022.
- [28] B. Di Martino, M. Rak, M. Ficco, A. Esposito, S. A. Maisto, and S. Nacchia, "Internet of things reference architectures, security and interoperability: A survey," *Internet of Things*, vol. 1, pp. 99–112, 2018.
- [29] D. Rahadiyan, S. Hartati, and A. P. Nugroho, "Design of an Intelligent Hydroponics System to Identify Macronutrient Deficiencies in Chili," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 1, 2022.
- [30] K. Reese, T. Smith, J. Dutton, J. Armknecht, J. Cameron, and K. Seamons, "A Usability Study of Five Two-Factor Authentication Methods," in *Fifteenth Symposium on Usable Privacy and Security (SOUPS 2019)*, 2019, pp. 357–370.
- [31] D. Yendri, "Two Sequential Authentication Method on Locker Security System Using Open-Sourced Smartphone," *JITCE (Journal Inf. Technol. Comput. Eng.)*, vol. 3, no. 02, pp. 65–69, 2019.
- [32] T. Alam, "Cloud-based IoT applications and their roles in smart cities," *Smart Cities*, vol. 4, no. 3, pp. 1196–1219, 2021.
- [33] O. Elijah, T. A. Rahman, I. Orikumhi, C. Y. Leow, and M. H. D. N. Hindia, "An overview of Internet of Things (IoT) and data analytics in agriculture: Benefits and challenges," *IEEE Internet Things J.*, vol. 5, no. 5, pp. 3758–3773, 2018.
- [34] W.J. Hong, N. Shamsuddin, E. Abas, R.A. Apong, Z. Masri, H. Suhaimi, S.H. Gödeke, and M.N.A Noh, "Water quality monitoring with arduino based sensors," *Environments*, vol. 8, no. 1, p. 6, 2021.
- [35] A. C. Bento, "IoT: NodeMCU 12e X Arduino Uno, Results of an experimental and comparative survey," *Int. J.*, vol. 6, no. 1, 2018.
- [36] T. Monica and A. S. Suneel, "Pollution Monitoring System Using Raspberry Pi and Php Web Server," in *Proceedings of the 2nd International Conference on Computational and Bio Engineering*, 2021, pp. 365–379.
- [37] G. Sai Pravallika, M. Lakshmi Akhila, M. Divya, T. Madhu Babu, and G. Kranthi Kumar, "Hand Gesture Controlled Contactless Elevator," in *Inventive Communication and Computational Technologies*, Springer, 2022, pp. 719–730.
- [38] M. U. H. Al Rasyid, M. H. Mubarrok, and J. A. N. Hasim, "Implementation of environmental monitoring based on KAA IoT platform," *Bull. Electr. Eng. Informatics*, vol. 9, no. 6, pp. 2578–2587, 2020.
- [39] [3M. Henschke, X. Wei, and X. Zhang, "Data visualization for wireless sensor networks using ThingsBoard," in *2020 29th Wireless and Optical Communications Conference (WOCC)*, 2020, pp. 1–6.
- [40] J. Rei, C. Brito, and A. Sousa, "Assessment of an IoT platform for data collection and analysis for medical sensors," in *2018 IEEE 4th International Conference on Collaboration and Internet Computing (CIC)*, 2018, pp. 405–411.
- [41] Y. Li, "An integrated platform for the internet of things based on an open source ecosystem," *Futur. Internet*, vol. 10, no. 11, p. 105, 2018.
- [42] "► Enterprise IoT Platform with Free Plan | Kaa," <https://www.kaaiot.com/> (accessed Apr. 06, 2022).
- [43] J. Yang, A. Sharma, and R. Kumar, "IoT-based framework for smart agriculture," *Int. J. Agric. Environ. Inf. Syst.*, vol. 12, no. 2, pp. 1–14, 2021.
- [44] B. E. Agossou and T. Toshiro, "IoT & AI Based System for Fish Farming: Case study of Benin," in *Proceedings of the Conference on Information Technology for Social Good*, 2021, pp. 259–264.
- [45] AshifuddinMondal and Z. Rehena, "IoT based intelligent agriculture field monitoring system," in *2018 8th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, 2018, pp. 625–629.



- [46] S. Vishwanath, S. Sharma, K. Deshpande, and S. Kanchan, "Vehicle Parking Management System," in 2020 International Conference on Convergence to Digital World-Quo Vadis (ICCDW), 2020, pp. 1–6.
- [47] "IoT Analytics - ThingSpeak Internet of Things." <https://thingspeak.com/> (accessed Apr. 06, 2022).
- [48] Á. Luis, P. Casares, J. J. Cuadrado-Gallego, and M. A. Patricio, "PSON: A Serialization Format for IoT Sensor Networks," *Sensors*, vol. 21, no. 13, p. 4559, 2021.
- [49] R. K. Kodali and V. S. K. Gorantla, "Restful motion detection and notification using iot," in 2018 International Conference on Computer Communication and Informatics (ICCCI), 2018, pp. 1–5.
- [50] L. O. Aghenta and M. T. Iqbal, "Low-cost, open source IoT-based SCADA system design using thinger. IO and ESP32 thing," *Electronics*, vol. 8, no. 8, p. 822, 2019.
- [51] "Thinger.io – Open Source IoT Platform." <https://thinger.io/> (accessed Apr. 06, 2022).
- [52] A. A. Ismail, H. S. Hamza, and A. M. Kotb, "Performance evaluation of open source IoT platforms," in 2018 IEEE global conference on internet of things (GCIoT), 2018, pp. 1–5.
- [53] L. T. De Paolis, V. De Luca, and R. Paiano, "Sensor data collection and analytics with thingsboard and spark streaming," in 2018 IEEE workshop on environmental, energy, and structural monitoring systems (EESMS), 2018, pp. 1–6.
- [54] S. A. Alavi, A. Rahimian, K. Mehran, and J. M. Ardestani, "An IoT-based data collection platform for situational awareness-centric microgrids," in 2018 IEEE Canadian conference on electrical & computer engineering (CCECE), 2018, pp. 1–4.
- [55] "ThingsBoard - Open-source IoT Platform." <https://thingsboard.io/> (accessed Apr. 06, 2022).

# Effective Multitier Network Model for MRI Brain Disease Prediction using Learning Approaches

N.Ravinder<sup>1</sup>

Research Scholar  
Department of CSE, Koneru Lakshmaiah Education  
Foundation, Vaddeswaram, AP, India

Dr. Moulana Mohammed<sup>2</sup>

Professor  
Department of CSE, Koneru Lakshmaiah Education  
Foundation, Vaddeswaram, AP, India

**Abstract**—Brain disease prognosis is considered a hot research topic where the researchers intend to predict the clinical measures of individuals using MRI data to evaluate the pathological stage and identifies the progression of the disease. With the lack of incomplete clinical scores, various existing learning-based approaches simply eradicate the score without ground truth score computation. It helps restrict the training data samples with robust and reliable models during the learning process. The major disadvantage of the prior approaches is the adoption of hand-crafted features, as these features are not well-suited for the prediction process. This research concentrates on modelling a weakly supervised multi-tier dense neural network model (*ws - MTDNN*) for examining the progression of brain disease using the available MRI data. The model helps analyze the incomplete clinical scores. The preliminary ties of the network model initially haul out the distinctive patches from the MRI to extract the global and local structural features (information) and develop a superior multi-tier dense neural network model for task-based image feature extraction and perform prediction in the successive tiers for computing the clinical measures. The loss function is adopted while examining the available individuals even in the absence of ground-truth values. The experimentation is done with the available online Dataset like ADNI-1/2, and the model works effectually with this Dataset compared to other approaches.

**Keywords**—Brain disease; learning approaches; ground truth value; feature learning; global and local feature analysis

## I. INTRODUCTION

Magnetic resonance imaging (MRI) is a suitable imaging technique for the head (specifically the brain) used in everyday clinical practice. It enables doctors to assess the nervous system's health and identify the existence of specific disorders. The computer-aided Alzheimer's disease (AD) prediction and premonitory phase, moderate cognitive decline (MCI), has made extensive use of MRI in recent years [1]–[6]. Anatomical MRI may detect aberrant brain structure and find imaging biomarkers for Alzheimer's disease (AD) in medical settings without radiation or other invasive procedures. Lately, assessing the state of disease and forecasting outcomes of AD and MCI progress employing baseline (BL) MRI information has been a popular issue.

Although numerous machine-learning approaches have already been developed for risk ratings utilizing BL MRI [11], a frequent difficulty of current systems is inadequately labelled information; participants may ignore ground-truth

diagnostic marks at specific time, amongst 805 participants inside AD Neural correlates Initiative-1 (ADNI-1) database [7]–[10], only 622 & 631 individuals had full CDR-SB & MMSE ratings 24 months following BL time, correspondingly. Earlier research simply discarded patients with incomplete clinical ratings owing to the sensitivity of reinforcement methods. Coupe [11] evaluated improvements of two clinical indicators from MRI utilizing 186 participants with comprehensive ground-truth clinical ratings from ADNI-1. Removing individuals with incomplete scores reduces the training dataset, decreasing the efficiency and resilience of estimation techniques. Furthermore, earlier machine-learning approaches often fed predetermined interpretations [e.g., image strength and tissue volume inside regions-of-interest] to ensuing prediction models, even though these characteristics may not be optimum for estimation techniques decreasing prognosis effectiveness.

The performance of deep learning methods has inspired various researches to use convolutional neural networks (CNN) to identify MRI characteristics for identifying certain diseases. Moreover, these techniques often fall inside the supervised learning method, making it impossible to train networks using people whose diagnostic ground-truth scores aren't full. Using all relevant poorly classified models (training participants with inadequate ground-truth scores at key time-points) is critical in Magnetic resonance brain illness diagnosis.

This research proposes the weakly supervised deep neural network (*ws - MTDNN*) for cerebral illness prediction utilizing BL MRI and partial clinical ratings at several time points. We define the following MR images and then identify multi-resolution image patchwork which relies on AD-based features. Finally, the deep CNN for forecasting of different clinical ratings at several points in time is created. This CNN has a novel weighting nonlinear function that enables the systems to learn sparsely labelled training data. Unlike prior MRI-based investigations, our suggested *was - MTDNN* technique can train models using all accessible individuals, even if some lack medical ratings at key periods[12][13]. Also, our anatomy landmark-based multi-resolution patch extraction procedure may address the issue of limited data by employing texture features instead of whole 3-D MR images as training examples.

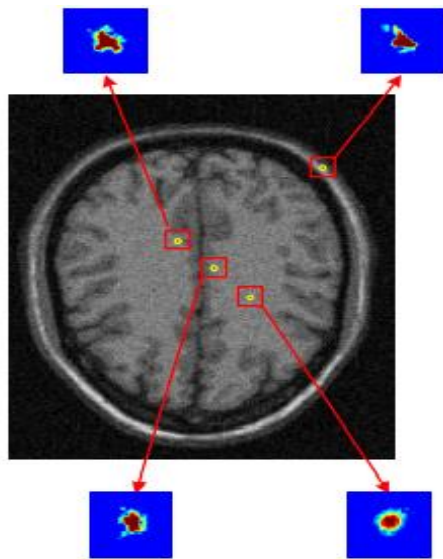


Fig. 1. Input MRI Image.

This section is used to summarise the paper's key contributions. First, unlike earlier researches [11], [17], we create a computational model with such a balanced gradient descent that can use all accessible weakly labelled patients (i.e., with partial ground-truth clinical ratings). It allows us to use all accessible labelled subjects. Integrating additional individuals in the training phase might aid in strengthening the resilience of the learnt system. Secondly, we suggest extracting variations in terms (as opposed to fixed-sized) input images if both small & large-scale patches centred at every location are retrieved. Our process is based on anatomic structures relevant to AD. This kind of approach assists in capturing the local/global analytical brain information [16]. Third, we create a combined prediction technique that simultaneously estimates many clinical scores at various times. The collaborative learning technique is anticipated to mimic the natural link between scores at/across various time points, aiding in the improvement of predictive performance. Using an MR image of a fresh experiment, the suggested technique can estimate four clinical ratings at four time-points in 12s, which is near to instant response. The research contributions are:

- 1) The input image is pre-processed for noise removal with Weiner filter and the contrast histogram equalization (CLAHE) is used for pixel block selection.
- 2) The weakly supervised multi-tier dense neural network model (  $ws - MTDNN$  ) is proposed to perform the classification process.
- 3) The performance is evaluated with indices like accuracy, specificity, sensitivity and error rate.

The work is provided as: Section II provides the comprehensive analysis of various prevailing approaches; Section III gives methodological analysis using pre-processing and IV is methodology explanation. The outcomes are discussed in Section 5 and summary in Section VI.

## II. RELATED WORK

In this part, we first discuss the standard interpretations of central nervous system MRIs before showcasing current MRI-based machine learning research to forecast and detect brain diseases. Many different features extracted from brain MRI have indeed been created in the research for automated AD/MCI prediction and diagnosis. These models may be loosely divided into three phases: voxel information, ROI recognition and patch recognition. More information on each type is provided as follows.

Voxel techniques [18] evaluate brain MRI by accurately measuring local tissues density (white and grey matter) following rigid normalization of actual brain images [19]-[20]. Sherubha et al. [21] suggested identifying volumetric information from specific brain parts from MR scans and then using them to categorize gender and AD. Moreover, voxel approaches are generally premised on the 1-1 anatomy mappings among participants and Gaussian dispersion of focal organ concentrations throughout statistical testing [23]. To suit the voxel description, tissue density is distorted with larger cones at the price of focused precision, which may lessen the voxel-based representation's racist and discriminatory potential for MRIs. Some downside of voxel-based modelling shows the amount of training data for individuals is typically quite small, resulting in the small-sample-size issue [24] and decreasing the effectiveness of learnt models. ROI-based depictions concentrate on assessing locally anatomical quantities in the mind's designated areas, in contrast to voxel-based characteristics. In example, earlier ROI-based studies often use tissue volume [11], [25]-[27], cortical thickness [28], hippocampal volume, and tissue densities in specific areas of the brain as feature extraction of MR data. This sort of representation, however, calls for an a priori hypothesis about aberrant areas from a structural standpoint to designate sections, even though these notions may not hold in actuality [22]. A defective brain area may cover numerous ROIs or only a tiny section of an ROI; therefore, employing a fixed brain division may reduce learning performance.

Patch-based analysis three was created to identify minute anatomical differences in brain MRIs using nonlinear analyses to represent the one-to-many mappings between brain structures. According to the author, the patch-based analysis may help diagnose AD and evaluate MCI development. Mohan et al. [28] used GM concentration within image regions as MRI for Disease prediction. The author suggested extracting morphometric information (local energy distribution) using AD-related anatomic structures. These carefully created MRI characteristics are often used to feed established models (such as SVMs and model structure [28]) for the diagnosis and prediction of diseases. However, given that the process of image retrieval and machine learning are carried out separately in these approaches, the pre-extracted MRI features in question may not be the most effective estimation techniques. Numerous supervised learning approaches have been developed to learn MRI characteristics that are task-oriented [14] [15]. An MR scan has millions of vertices, so many brain areas may not have been impacted by Alzheimer's. As a result, one of the main challenges in MRI-

based transfer learning is figuring out how to identify correctly (e.g., discriminatively across groups) in MRIs.

To overcome this trouble, Myronenko et al. [29] suggested concentrating on three ROIs (i.e., the hippocampus, ventricle, and neuroimaging surface) and created the deep CNN for risk that exists in measurements of participants using 2D texture features taken from the three ROIs. In brain scans (i.e., architectural MRI and mobility tensor image information), the CNN used the hippocampus ROI and adjacent areas. Similarly, the author published a deep ranking algorithm for classifying AD from the hippocampus ROI. These studies employ experimentally identified MRI areas without addressing other possibly essential brain regions. The author created a 2D CNN to distinguish AD patients using functional and structural MRI scans. However, they reduce 3D and 4D images to 2D slices and give inputs to the networks, neglecting the crucial spatial information. Recently, Risk et al. [30] developed an anatomic heritage site deep learning system for Clinical examination and MCI conversion prediction. To be more precise, they first identify 3-D image patches using brain regions with AD-related anatomic structures, and afterwards, they create a CNN for combined MRI extracting the features and disease categorization. However, set the size of texture features is employed in these investigations, disregarding the possibility that structural alterations brought on by dementia might differ significantly across various brain areas.

Additionally, most current deep learning techniques are completely regulated, with individuals lacking ground-truth ratings at certain intervals simply being eliminated. To properly engage all available patients (including those lacking ground-truth ratings) for training, a semi-supervised CNN is presented for prediction of MRI data. The suggested approach departs from the earlier research in [30]. In this research, we employ weakly labelled training items by designing a distinctive weighted nonlinear function in the suggested neural net, while earlier techniques [30] can only use completely labelled (whole ground-truth score) training cases. This article attempts to extract multi-resolution input images centred at each landmark site to simulate brain MRI multi-resolution spatial features, whereas only uses fixed-sized input images.

### III. DATA ACQUISITION

We conducted trials on 1469 individuals drawn from subsets of the accessible Dataset collected [10], namely ADNI-1/2. 805 participants have BL structured MRI data from ADNI-1, and 664 individuals from ADNI-2. Individuals are immediately deleted from ADNI-2 if they feature for both ADNI-1 and ADNI-2. In contrast to the participants in ADNI-1, who had 1.5 T T1-weighted MRI, ADNI-2 had 3.0 T T1-weighted MRI. In our investigations, ADNI-1/2 are two separate databases. These issues may be divided into three groups based on several criteria: AD, MCI, and HC.

Four clinical criteria are utilized: 1) CDR-SB; 2) ADAS-Cog11, a different form of the ADAS-Cog with 11 items; 3) ADAS-Cog13, a 13-item version of the ADAS-Cog; and 4) MMSE. The BL time following approval is the day individuals were supposed to conduct the screening. Additionally, the length beginning from the BL time indicates

the time points for obeying visits. Every participant under investigation has MRI data at baseline. However, many lack ground truth scores for certain clinical parameters at particular periods. Table I displays comprehensive details on the topics under study. For each subject's structural MR imaging, we first correct the anterior-posterior commissures, strip the skull and remove the cerebellar. Next, we align every image to a shared Colin27 template before resampling all MR images with a horizontal spatial resolution. Finally, we adjust brightness heterogeneities for each MR image using the N3 method.

#### A. Weiner Filter

It provides a substantial role in various applications like echo cancellation, linear prediction, signal restoration, system prediction and channel equalization. The Weiner coefficients are evaluated to reduce the average squared distance among the desired input and the filtered output. The proposed filter theory considers the inputs that are stationary process. When the filtering coefficients are re-evaluated at periodic intervals for every blocks of 'N' signal samples then the filter needs to adapt the average signal characteristics within the block and works block adaptively. It is determined to be stationary over the relatively small sample blocks. The target of Weiner filter is reducing the mean square error value and image restoration. It is expressed as in Eq. (1):

$$x(n) = d(n) + v(n) \quad (1)$$

Here,  $d(n)$  and  $v(n)$  represents stationary random process.

#### B. Contrast Limited Adaptive Histogram Equalization

The following are the CLAHE procedure:

1) Partition the image into number of equal sub-blocks (size) and every sub-block should be non-overlapping and continuous.

2) Measure the local histogram of every sub-block;

3) Evaluate the average number of pixels allocated to sub-block gray level ( $Avnum$ ). When  $GrayNum$  is utilized to specify the probable gray level to sub-blocks, the process is depicted in Eq. (2) where  $XP$  and  $YP$  represents the number of pixels in  $X$  and  $Y$  sub-block directions.

$$AvNum = \frac{XP.YP}{GrayNum} \quad (2)$$

4) The shear coefficient  $CV$  is fixed with a range of [0,1]. For various images, it can be adjusted to provide superior value via the simulation outcomes, and the actual shear limit value  $NV$  is expressed as in Eq. (3). Here, round specifies the rounding off function.

$$NV = Avnum + \text{round}(CV.(XP.YP - AvNum)) \quad (3)$$

5) With the shear limit, the pixels for every gray level of local histogram and the added number of pixels are re-distributed to every gray level of histogram. Consider, that  $Nclip$  specifies the total amount of pixels that are eliminated. Therefore, the number of pixels can be attained  $NA_{cp}$  that every gray level is allocated with Eq. (4) and Eq. (5):

$$Nclip = \sum(\max(H(i) - NV, 0)) \quad (4)$$

$$NA_{cp} = \frac{Nclip}{GrayNum} \quad (5)$$

Here,  $CH$  is the histogram after the re-distribution process and it is attained by Eq. (6):

$$CH(i) = \begin{cases} NV & H(i) > NV \\ NV & H(i) + NA_{cp} \geq NV \\ H(i) + NA_{cp} & else \end{cases} \quad (6)$$

6) Consider that the remaining amount of pixels after distribution is  $NumLeft$ , step distribution size is depicted as:

$$Step = \frac{GrayNum}{Numleft} \quad (7)$$

Initiate searching from minimal gray level by step size; therefore the pixels are allocated when the numbers of pixels are lesser than shear threshold. Then, finish the cycle from the minimal to maximal gray level till  $Numleft$  is set to zero. Therefore, the histogram allocated is fulfilled and some new histogram is acquired.

7) Histogram equalization is done on every sub-region after the shearing process.

8) The centre-point sub-blocks is considered after the reference point acquired from the gray value. Every image pixel is executed by bilinear interpolation and pixel mapping is provided using the related regions with adjacent reference points. Assume, the small rectangle  $(x, y)$  specifies target point and  $f(x, y)$  represents gray value to evaluate  $(x, y)$ . The adjacent regions' center point is provided as  $A(x-, y-), B(x+, y-), C(x-, y+)$  and  $D(x+, y+)$ . The gray level value  $f(x, y)$  is specified as linear gray value combination with four points. For every pixel over the boundary regions, gray level is evaluated using the linear interpolation adjacent sample points where the corner points are evaluated with the adjacent sample points as in Eq. (8):

$$f(x, y) = a[bf(x-, y-) + (1 - b)f(x+, y-)] + (1 - a)[bf(x-, y+) + (1 - b)f(x+, y+)] \quad (8)$$

$$a = \frac{y-(y-)}{(y+)-(y-)} \quad (9)$$

$$b = \frac{x-(x-)}{(x+)-(x-)} \quad (10)$$

#### IV. METHODOLOGY

In this study, we aim to address two difficult issues in MRI-based brain illness prediction: how to fully exploit poorly labelled training data (i.e., individuals with inadequate ground-truth medical ratings) and how to learn important characteristics of MR images structurally. A weakly supervised CNN is created to incorporate extraction of features and model learning into a cohesive framework, using all accessible weakly labelled subjects for the training phase. The suggested  $ws - MTDNN$  approach consists of two basic steps: extracting multi-scale image patches and classification. More information is provided below. There are thousands of voxels for each brain MR imaging, yet dementia's structural alterations may be minor. When the complete MR image is

provided to the deep learning model, the inputs contain much loud noise data, making network development challenging with just a few hundred training subjects. To train the classifier for accurate illness prediction, we want to find important brain areas in each MRI rather than utilizing the complete image.

We use anatomical markers to find AD-related areas. Here, landmark detection is used to derive 1741 anatomical structures from the Colin27 templates. Numerous landmarks are geographically adjacent to one another, as can be seen in the supplemental materials in Fig. 2. To avoid data duplication and computation time, we chose  $K = 40$  anatomical landmarks. We initially sorted those features in order of increase using the p-values that the landmark identification method obtained via a correlation among AD and HC individuals. The spatial Distance measure is provided to limit (i.e., 20) the separation among landmarks, and we use the top  $K = 40$  monuments. For example, we display the recognized features on three individuals and these landmarks in the templates region.

We generate multi-resolution texture features from an input MRI using these landmarks to obtain extra depth information. Specifically, we derive both small and large scale regions out of MRI. Those patches are all centred on the respective landmarks. So, provide  $K$  landmarks, we can get  $2K$  image patches of certain MRI. These multi-resolution image patching serve as the input information for suggested model.

We simultaneously conduct pattern recognition of MRIs and recovery of numerous clinical ratings at four consecutive using multi-scale image patches from each MRI using the suggested human brain. The proposed scheme receives  $2K$  image patches from each participant as inputs, and it outputs four different clinical measurements at 4 various time points.

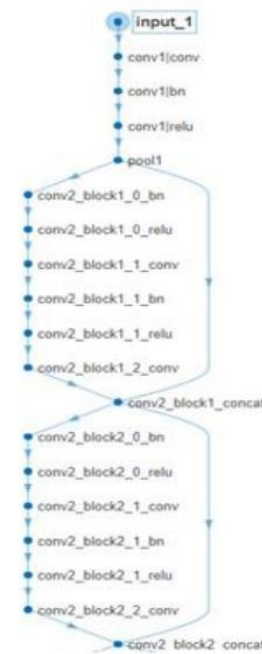


Fig. 2. Proposed Model Network Architecture.



We initially concentrate on modelling the nearby spatial features seen in multi-scale image patchwork using  $K$  parallel subnets, each mapped to a particular landmark point. Within every sub-network, the investigators first decrease the effect of the major large-scale update. Therefore, it is a relatively similar diameter to the small-scale patch. These tiny patches are therefore individually sent into a sub-network consisting of three deep convolutional modules (DCMs) and two fully connected (FC) levels, treating them as the two-channel input. In each DCM, three convolutional neural networks are followed by a 222 max-pooling plane for output feature wavelet decomposition. In distinctive, for a precise convolution operation within every DCM, the feature maps (the images that come out of each convolution operation) of all the layers before they are being used as inputs, and the convolution layers of all the layers after they have been used as inputs. Batch standardization and linear transfer unit (ReLU) activation are used after each convolution operation. Such densely linked design strengthens feature propagation, encourages feature reuse, and reduces network parameter optimization. The  $K$  parallel subnetworks have identical designs but individually optimized characteristic weights. We want to discover landmark local characteristics from images using  $K$  sub-networks to preserve each landmark site's distinctive local analytical information. If sub-networks share properties, we can't extract historic site-local spatial features from brain MRIs.

It is important to note that the overall architecture of an MRI cannot be captured by utilizing merely the local patches alone. To do this, the feature maps knowledge gained since the last  $K$  FC layers in  $K$  sub-networks are added together, and then two more FC layers are added to learn the neighbourhood classification model of the information MR image. Four clinical-grade categories at four different time points are predicted using the last FC layer (containing 32 neurons). Based on [3], we created a weighted loss function for the network model so that all available loosely labelled training participants could be used to their fullest (missing ground-truth clinical scores). We will refer to the training set of  $N$  individuals as  $X = [x_1, \dots, x_n, \dots, x_N]$ , where  $W$  refers to the network coefficients. Its  $s$ th ( $s = 1, \dots, S$ ) ground-truth clinical value at the  $t$ -th ( $t = 1, \dots, T$ ) time-point is indicated as  $y_n^{s,t}$  for the  $n$ th ( $n = 1, \dots, N$ ) subject  $x_n$ . The suggested optimization problem aims to reduce the gap here between the following projected number  $f^{s,t}(x_n; W)$  and the actual number  $y_n^{s,t}$ :

$$\arg \min_W \frac{1}{N} \sum_{n=1}^N \frac{1}{\sum_{s=1}^S \sum_{t=1}^T \gamma_n^{s,t}} \sum_{t=1}^T \gamma_n^{s,t} * \left( y_n^{s,t} - f^{s,t}(x_n; W) \right)^2 \quad (11)$$

Where  $\gamma_n^{s,t}$  indicates whether or not  $x_n$  is given the  $s$ th medical value at the  $t$ -th time-point. In particular, if the ground-truth score  $y_n^{s,t}$  is accessible for  $x_n$ , then  $\gamma_n^{s,t} = 1$ . To be more particular, even if an instructional subject has omitted scores at some points in time and therefore does not start contributing to the loss of data processing (i.e.,  $\gamma_n^{s,t} = 0$ ), it still start contributing to the logistic regression during network training. Therefore, increased throughput is used at various time points. Furthermore, we may have used all accessible individuals (even if they lack ground reality diagnostic ratings

at various time intervals) for model training using Eq. (11). It seems possible because (1) allows us to build representations from MR scans informally automatically. The typical beginning of the module dismisses individuals with insufficient ground-truth scores, in contrast to this.

We randomly choose alternative patches centered at each landmark position with separations, and the phase margin is one. This one will increase the training set and lessen the detrimental effect of landmark identification mistakes. As a result, each MRI may also provide 125 patches, one for each point, at each scale. Given  $K$  landmarks, we may create 125K variations of patched at each size, each serving as a unique sampling for the neural framework. It technically allows us to create 125K examples for MRI, but these sampling are utilized as input information randomly for the suggested system.

At the training step, designers use the instructional subject matters' BL MRIs as inputs and their own ground-truth diagnostic and therapeutic goals scored at four points in time (with incomplete data) as outputs to train the network. In particular, we firstly collect variations in terms (i.e., 242424 and 484848) image regions from each train MRI and then input such patched to the networks  $k$  – Means and  $k$  anatomic structures. This method may discover a mapping function from every MRI source to the three clinical ratings at four different periods. During testing step, we first identify its related landmarks using deep learning for an unknown experiment with just a BL MR image and then create a multi-resolution patchwork. We next input these multi-resolution image patches to the trained network to forecast the clinical ratings at four different periods for this test patient.

Stochastic gradient descent (SGD) and the back-propagation technique for generating network concentrations and updating parameters are used to improve the objective function. The mobility parameter and the number of iterations for SGD were explicitly calibrated experimentally to 0.9 and 104. Fig. 3 displays the changing curves for the calibration and testing loss functions on the ADNI-1 database. Using a computer with a solitary GPU, our process utilizes about 12 seconds to forecast the four types of diagnostic tests of the MRI experiment. Inferring the suggested  $ws$  – *MTDNN* approach is anticipated to carry out real-time brain illness diagnosis in practical applications. The software and primarily targeted model are accessible online.

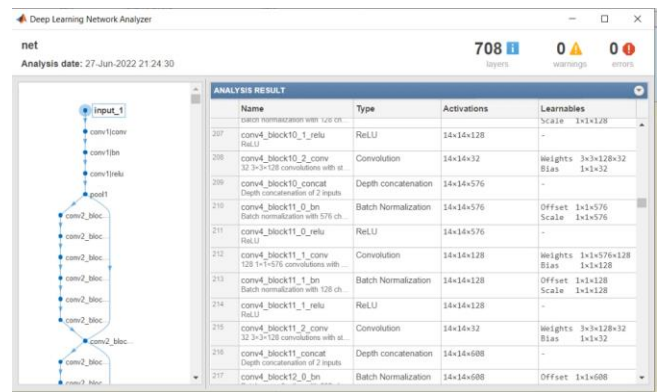


Fig. 3. Layer Description.



### V. NUMERICAL RESULTS AND ANALYSIS

We execute two sets of trials in twofold confidence intervals to test the suggested method's resilience. We explicitly train models on participants from ADNI-1 and evaluate them from the separate ADNI-2 Dataset in the first set of trials. The second category trains on ADNI-2 and tests on ADNI-1. Various performance metrics like accuracy, specificity, sensitivity and error rate are evaluated and compared with other approaches. The expressions for these metrics are given below:

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (12)$$

$$Sensitivity = \frac{TP}{TP+FN} \quad (13)$$

$$Specificity = \frac{TN}{TN+FP} \quad (14)$$

$$Error\ rate = 1 - \frac{1}{2} (sensitivity + specificity) \quad (15)$$

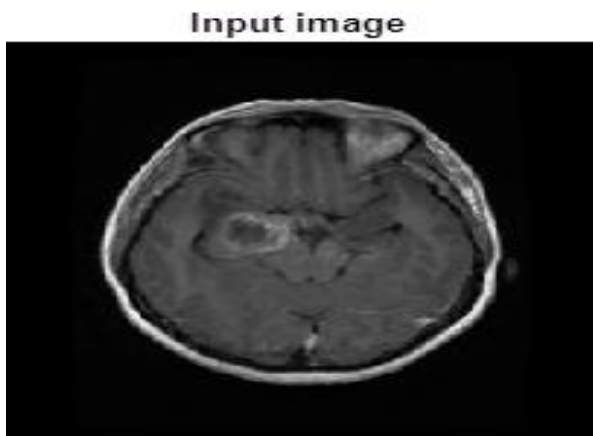


Fig. 4. Input Image.

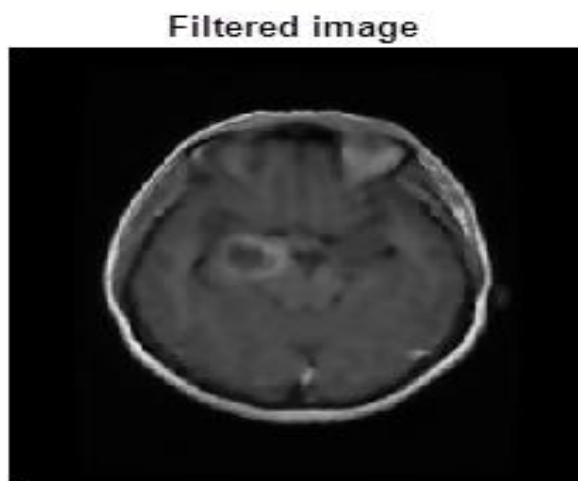


Fig. 5. Filtered Image.

Fig. 4 to Fig. 6 provides the outcome of the pre-processed image. Table 1 compares approaches like conventional ANN, SVM, BoVW-based SVM, conventional CNN and DNN. Here, metrics like accuracy, specificity, sensitivity and error rate are evaluated and compared with other approaches. The

accuracy of the anticipated model is 93.08% which is 58.08%, 2.08%, 1.08%, 20.08% and 25.08% higher than other approaches. The specificity of the anticipated model is 83.47% which is 48.47%, 7.47% and 15.47% higher than ANN, conventional CNN and DNN and 6.53% and 9.53% lesser than SVM and BoVW-based SVM model. The specificity of the anticipated model is 100% which is 65%, 9%, 7%, 31% and 35% higher than other methods. The error rate is 0.069 for the anticipated model, which is comparatively lesser than other approaches. Other approaches pose an error rate of 1.2568, 2.564, 3.548, 1.565 and 16.235, respectively. Based on the analysis, it is proven that the anticipated model works well compared to other approaches in the prediction process (See Fig. 7 to Fig. 10).



Fig. 6. Equalized Image.

TABLE I. OVERALL COMPARISON

S. No	Methods	Accuracy	Sensitivity	Specificity	Error rate
1	ANN	35	35	35	1.2568
2	SVM	91	90	91	2.564
3	VW-based SVM	92	93	93	3.548
4	Conventional CNN	73	76	69	1.565
5	DNN	68	68	65	16.235
6	WS-MTDNN	93.08	83.47	100	0.069

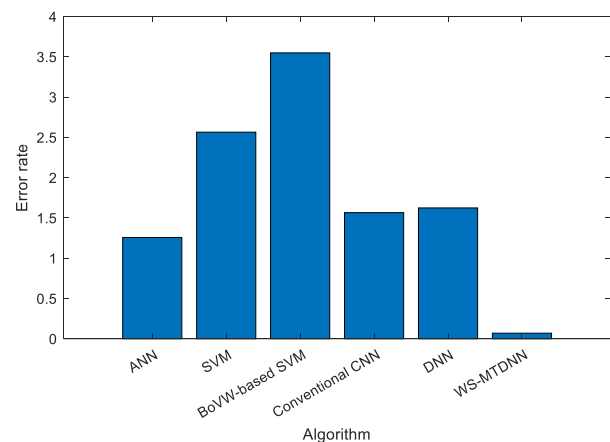


Fig. 7. Error Rate Comparison.

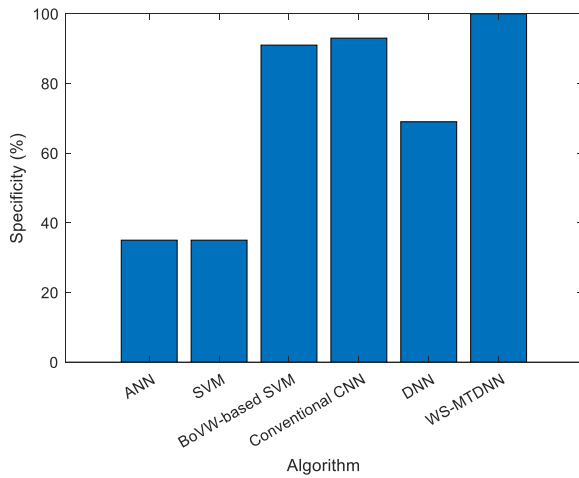


Fig. 8. Specificity Comparison.

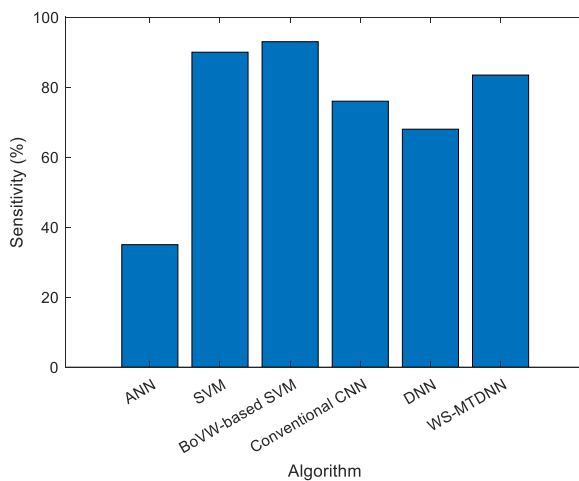


Fig. 9. Sensitivity Comparison.

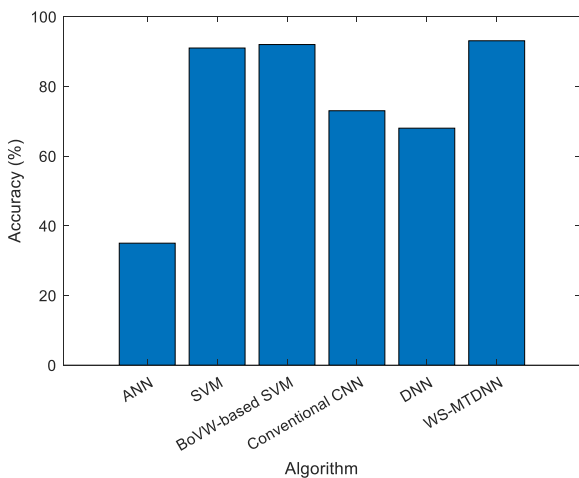


Fig. 10. Accuracy Comparison.

#### A. Constraint Analysis

The following is a summary of the restrictions this paper still has.

1) The suggested technique was only evaluated for predicting clinical values from MRI images, but ADNI-1 and ADNI-2 databases include transverse MRI scans. The issue with utilizing longitudinally MRI scans for therapeutic score prediction lack subsequent images.

2) The process largely functions for estimating various clinical scores while measuring the relationship among the clinical scores and subject classifications (such as AD or HC).

3) Local patch identification based on anatomic structures is autonomous of extracting the features and classification building, which may hinder prognosis performance.

4) We did not consider the differences in the subject distribution of data between ADNI-1 and ADNI-2. It might adversely impact the generalizability of our technique.

As a result:

1) Users will translational MRI scans to assess the clinical scores. For accurate prediction, missing MRI scans will be filled with learning algorithms (like generative adversarial). Also, full (after calculation) MRI images for evaluating clinical grades at all time points may indicate which time point would be most relevant in disease progression.

2) Given the strong correlation between clinical values and membership functions for a given patient, it seems sensible to create a single deep learning model that combines analysis and categorization.

3) Immediately detect patch/region-level racially discriminatory spots in the entire MRI so patch and region organization contains may be concurrently learnt and merged to build illness classification techniques.

4) We want to develop a better classification approach to address the issue of diverse data distributions. It is anticipated to better the suggested network's capacity to generalize.

#### VI. CONCLUSION

In this research, we suggested ws – MTDNN for predicting many clinical scores based on individuals having MRI data and partial clinical ratings. It was done using individuals as training data and individuals as validation data. Specifically, we pre-processed all MR images and then used feature detection algorithms to locate disease-related anatomical structures in the patients' bodies. Based on the position of each landmark, we selected multi-scale patchwork with the landmarks serving as their centres. We constructed a deep convolutional neural network to concurrently learn discriminant information from MRI and forecast several clinical grades at four different periods. The input data for this network were image patches. Our network model constructed a balanced loss function for all training patients, though many may not have full ground-truth clinical ratings. The suggested ws – MTDNN algorithm can identify clinical grades at future time points utilizing MRI data in science experiments from the available datasets.

#### REFERENCES

- [1] Abiwinanda N, Hanif M, Hesaputra ST, Handayani A, Mengko TR (2019) Brain tumour classification using convolutional neural network.

- In: World congress on medical physics and biomedical engineering 2018. Springer, pp 183–189.
- [2] Al-Zu'bi S, Hawashin B, Mughaid A, Baker T (2021) Efficient 3d medical image segmentation algorithm over a secured multimedia network. *Multimed Tools Appl* 80(11):16887–16905.
- [3] Ahmed KB, Hall LO, Goldof DB, Liu R, Gatenby RA (2017) Fine-tuning convolutional deep features for MRI-based brain tumour classification. In: *Medical imaging 2017: computer-aided diagnosis*, vol 10134. international society for optics and photonics, p 101342e.
- [4] Alzu'bi S, Jararweh Y, Al-Zoubi H, Elbes M, Kanan T, Gupta B (2019) Multi-orientation geometric medical volumes segmentation using 3d multi-resolution analysis. *Multimed Tools Appl* 78(17):24223–24248.
- [5] AlZu'bi S, Al-Qatawneh S, Alsmirat M (2018) Transferable hmm trained matrices for accelerating statistical segmentation time. In: *2018 Fifth international conference on social networks analysis, management and security (SNAMS)*. IEEE, pp 172–176.
- [6] Baldi P (2012) Autoencoders, unsupervised learning, and deep architectures. In: *Proceedings of ICML Workshop on unsupervised and transfer learning*, pp 37–49.
- [7] AlZu'bi S, Shehab M, Al-Ayyoub M, Jararweh Y, Gupta B (2020) Parallel implementation for 3d medical volume fuzzy segmentation.
- [8] Greenspan, "Super-resolution in medical imaging," *Comput. J.*, vol. 52, no. 1, pp. 43–63, 2008.
- [9] Manjón, P. Coupé, A. Buades, V. Fonov, D. L. Collins, and M. Robles, "Non-local MRI upsampling," *Med. Image Anal.*, vol. 14, no. 6, pp. 784–792, Dec. 2010. Manjón, P. Coupé, A. Buades, D. L. Collins, and M. Robles, "MRI super-resolution using self-similarity and image priors," *Int. J. Biomed. Imag.*, vol. 2010, Dec. 2010, Art. No. 425891.
- [10] Coupé, J. V. Manjón, M. Chamberland, M. Descoteaux, and B. Hiba, "Collaborative patch-based super-resolution for diffusionweighted images," *NeuroImage*, vol. 83, pp. 245–261, Dec. 2011.
- [11] Jafari-Khouzani, "MRI upsampling using feature-based non-local means approach," *IEEE Trans. Med. Imag.*, vol. 33, no. 10, pp. 1969–1985, Oct. 2014.
- [12] Hu, X. Wu, and J. Zhou, "Second-order regression-based MR image upsampling," *Comput. Math. Methods Med.*, vol. 2017, Jan. 2017, Art. No. 6462832.
- [13] Buades, B. Coll, and J.-M. Morel, "A non-local algorithm for image denoising," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2, Jun. 2005, pp. 60–65.
- [14] Takeda, S. Farsiu, and P. Milanfar, "Deblurring using regularized locally adaptive kernel regression," *IEEE Trans. Image Process.*, vol. 17, no. 4, pp. 550–563, Apr. 2008.
- [15] Wang, "Multi-scale image sharpening adaptive to edge profile," *J. Electron. Imag.*, vol. 14, no. 1, Jan. 2005, Art. no. 013.
- [16] Lopez-Rubio, "Superresolution from a single noisy image by the median filter transform," *SIAM J. Imag. Sci.*, vol. 9, no. 1, pp. 82–115, 2016.
- [17] Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 2, pp. 295–307, Feb. 2016.
- [18] Kim, J. K. Lee, and K. M. Lee, "Accurate image super-resolution using intense convolutional networks," in *Proc. CVPR*, Jun. 2016, pp. 1646–1654.
- [19] Sherubha, "Graph Based Event Measurement for Analyzing Distributed Anomalies in Sensor Networks", *Sādhanā(Springer)*, 45:212, <https://doi.org/10.1007/s12046-020-01451-w>.
- [20] Sherubha, "An Efficient Network Threat Detection and Classification Method using ANP-MVPS Algorithm in Wireless Sensor Networks", *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, ISSN: 2278-3075, Volume-8 Issue-11, September 2019.
- [21] Sherubha, "An Efficient Intrusion Detection and Authentication Mechanism for Detecting Clone Attack in Wireless Sensor Networks", *Journal of Advanced Research in Dynamical and Control Systems (JARDCS)*, Volume 11, issue 5, Pg No. 55-68.
- [22] Elbes M, Alzubi S, Kanan T, Al-Fuqaha A, Hawashin B (2019) A survey on particle swarm optimization with an emphasis on engineering and network applications. *Evol Intel* 12(2):113–129.
- [23] Guo X, Yin Y, Dong C, Yang G, Zhou G (2008) On the class imbalance problem. In: *2008 Fourth international conference on natural computation*, vol 4. IEEE, pp 192–201.
- [24] Gumaie A, Hassan MM, Hassan MR, Alelaiwi A, Fortino G (2019) A hybrid feature extraction method with regularized extreme learning machine for brain tumour classification. *IEEE Access* 7:36266–36273.
- [25] Jain R, Jain N, Aggarwal A, Hemanth DJ (2019) Convolutional neural network-based Alzheimer's disease classification from magnetic resonance brain images. *Cogn Syst Res* 57:147–159.
- [26] Kumar S, Dabas C, Godara S (2017) Classification of brain MRI tumour images: a hybrid approach. *Procedia Comput Sci* 122:510–517.
- [27] Mohan G, Subashini MM (2018) Mri based medical image analysis: Survey on brain tumour grade classification. *Biomedical Signal Processing and Control* 39:139–161.
- [28] Myronenko A (2018) 3d MRI brain tumour segmentation using autoencoder regularization. In: *International MICCAI brain lesion workshop*. Springer, pp 311–320.
- [29] Rizk H, Shokry A, Youssef M (2019) Effectiveness of data augmentation in cellular-based localization using deep learning. [arXiv:1906.08171](https://arxiv.org/abs/1906.08171).

# Application based on Hybrid CNN-SVM and PCA-SVM Approaches for Classification of Cocoa Beans

AYIKPA Kacoutchy Jean<sup>1</sup>, MAMADOU Diarra<sup>2</sup>, BALLO Abou Bakary<sup>3</sup>, GOUTON Pierre<sup>4</sup>, ADOU Kablan Jérôme<sup>5</sup>

ImVia, Université Bourgogne Franche-Comté, Dijon, FRANCE<sup>1, 4</sup>

LaMI, Université Felix Houphouët-Boigny, Abidjan, CÔTE D'IVOIRE<sup>1, 2, 3, 5</sup>

UREN, Université Virtuelle de Côte d'ivoire, Abidjan, CÔTE D'IVOIRE<sup>1</sup>

**Abstract**—In our study, we propose a hybrid Convolutional Neural Network with Support Vector Machine (CNN-SVM) and Principal Component Analysis with support vector machine (PCA-SVM) methods for the classification of cocoa beans obtained by the fermentation of beans collected from cocoa pods after harvest. We also use a convolutional neural network (CNN) and support vector machine (SVM) for the classification operation. In the case of the hybrid model, we use a convolutional network as a feature extractor and the SVM is used to perform the classification operation. The use of PCA-SVM allowed for a reduction in image size while maintaining the main features still using the SVM classifier. Radial, linear and polynomial basis function kernels were used with various control parameters for the SVM, and optimizers such as the Stochastic Gradient Descent (SGD) algorithm, Adam, and RMSprop were used for the CNN softmax classifier. The results showed the robustness of the hybrid CNN-SVM model which obtained the best score with a value of 98.32% then the PCA-SVM based model had a score of 97.65% outperforming the standard CNN and SVM classification algorithms. Metrics such as accuracy, recall, F1 score, mean squared error (MSE), and MCC have allowed us to consolidate the results obtained from our different experiments.

**Keywords**—Support vector machine; convolutional neural network; cocoa beans; principal component analysis; hybrid method

## I. INTRODUCTION

Côte d'Ivoire is the world's largest producer of cocoa [1] and cocoa is an important cash crop in the world. The cocoa culture produces beans from the ripe pod seeds of the Theobroma plant [2]. Cocoa beans are the main raw material for chocolate [3] and the first step in this process is fermentation.

Fermentation is an essential step in cocoa processing, and it has an impact on the flavor, color, and aroma of cocoa products [4]. Unfermented cocoa beans do not have the full flavor of chocolate, but fermentation triggers chemical changes within the cocoa bean that contribute to the development of chocolate flavor [5]. Once harvested, farmers open the cocoa pods, extract the cocoa seeds with the pulp and fill wooden boxes or containers to begin fermentation [6]. Using quality dried cocoa beans from the fermentation process allows for obtaining better-finished products. The process of detecting the quality of dried cocoa beans is a tedious task and requires special attention, hence the need to use computer vision that will allow an image to specify its category of it. In recent years, computer vision has an important role in agricultural production with the

use of machine learning and more specifically deep learning convolutional neural networks [7]. We can note here the popular image analysis techniques in machine learning such as Support vector machine (SVM), Artificial Neural Network (ANN), Convolutional Neural Networks (CNN), Normalized Difference Vegetation Index (NDVI), and statistical tools such as correlation and regression analysis, etc [8]. In order to facilitate the classification of cocoa beans, we proposed the methods of CNN, SVM, hybrid CNN-SVM, and principal component analysis with support vector machine (PCA-SVM). The general principle of the hybrid model is to automatically extract features based on the CNN and do the classification using the SVM classifier, while the PCA-SVM reduces the image size while keeping the main features still using the SVM classifier. All these methods were used to detect the category of cocoa dried seeds and then we compared them to come out with the best method. Our technique can evolve into an industrial application with an appropriate integration framework, replacing the traditional method of quality control of cocoa beans. Thus, our study can be integrated into a computer vision system and implemented in the cocoa production and processing chain, resulting in a state-of-the-art automatic solution. The proposed approach could benefit the industry by enabling them to accurately determine the quality of cocoa beans.

This paper is organized as follows: Section II presents some previous work, and Section III details the methodology and the material we propose. Section IV presents the results obtained and discusses them, and we conclude in Section V.

## II. RELATED WORK

Several works focused in this area and we can cite some of them, namely, Oliviera et al. used handcrafted features calculated from the beans provided by image analysis tools, and then the random decision forest predictor was used to classify the samples. This experiment yielded an accuracy of 92% [9]; Kaghi et al. used a pre-trained AlexNet CNN as a generic feature extractor of a 2D image whose dimensions were reduced using PCA + TSNE and finally classified using a simple machine learning algorithm like KNN, and Naïve Bayes Classifier. These results could match a CNN Softmax classifier [10]; Barbon et al. used machine learning-based methods namely J48, Naïve Bayes, K-NN, Random Forest, SVM, MLP, and Fuzzy approaches to predict the storage time of pork, and these methods provided the accuracies which ranged from 78.26 to 94.41% [11]. A. and Renjith provide a special

architecture to identify the features of different classes of Durian fruit, the image processing model allows the classification to be divided into two parts: feature extraction and classification of the fruit used. The use of edge detection and color extraction provide correct feature extraction of durian, the performance is measured using non-destructive machine learning techniques such as SVM, GNB, and Random Forest. The results obtained provide the best accuracy of 89.3% using the SVM technique and 84.3% using Random Forest [12]; Harel et al. proposed maturity classification algorithms. Their algorithms were applied to the maturity classification of peppers. Maturity classification achieved 98.2% and 97.3% accuracy for two-class classification between mature and immature classes of red and yellow peppers, respectively, and 89.5% and 97.3% accuracy for four-class maturity classification. The random forest algorithm has been shown to be very robust [13]. Elleuch et al. explored a method focusing on the use of two classifiers in this case Convolutional Neural Network (CNN) and Support Vector Machine (SVM) for offline Arabic handwriting recognition, the performance of their methods was compared with the character recognition accuracies obtained from the state of the art of optical Arabic character recognition, producing favorable results [14].

### III. MATERIAL AND METHOD

#### A. Materials

The digital images of cocoa beans used for the study were obtained at YAKASSE 1 (longitude: -3.77374 latitude: 5.23841) village located near GRAND-BASSAM in Côte d'Ivoire. These cocoa bean samples were classified as follows:

- Category 1 beans: fermented and dried cocoa beans of superior quality.
- Category 2 beans: fermented and dried cocoa beans of intermediate quality.
- Category 3 beans: non-fermented and dried cocoa beans.

Once the data is obtained as shown in Fig. 1, we will proceed to the pre-processing that will extract the seeds from each image.

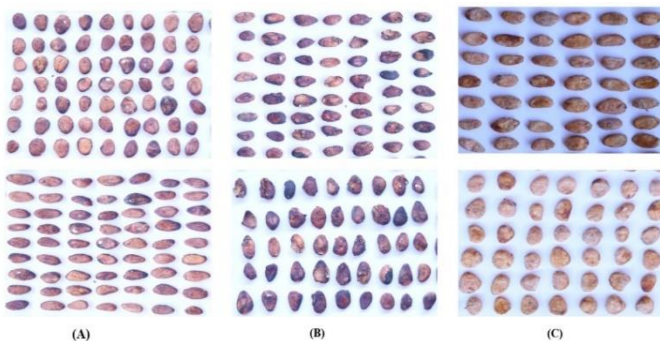


Fig. 1. Cocoa Beans Sample Images from the Three Classes: (A) Category 1 Beans; (B) Category 2 Beans; (C) Category 3 Beans.

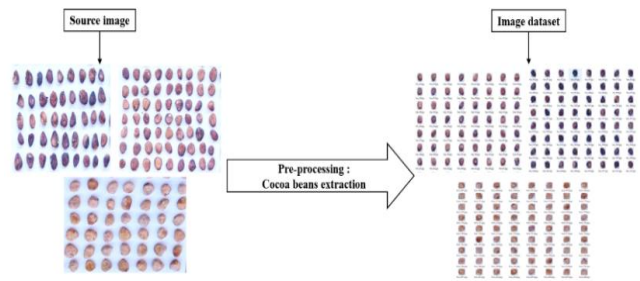


Fig. 2. General Schema of the Dataset.

We describe the preprocessing stage now

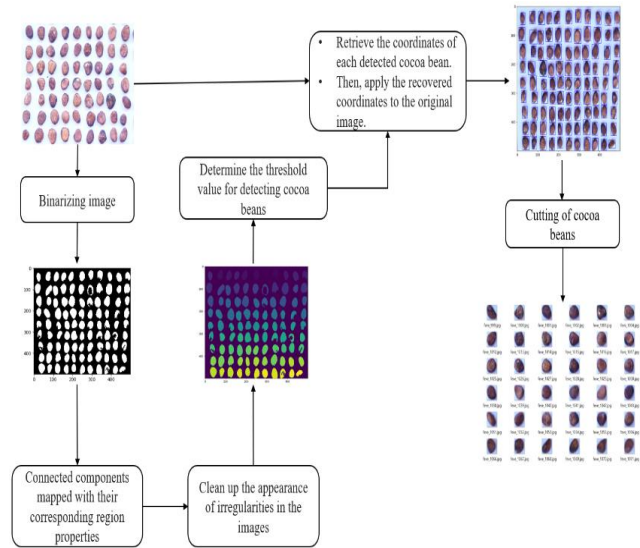


Fig. 3. Preprocessing Cocoa Beans Extraction.

After the preprocessing step as shown in Fig. 2 and Fig. 3, we obtained a dataset of 3470 images of cocoa beans, including 917 images of beans of category 1, 1675 images of beans of category 2, and 878 images of beans of category 3. We split the dataset into 60% for training, 20% for testing, and 20% for validation.

The table I presents the data split description-

TABLE I. DESCRIPTION OF THE DATA SPLIT

Dataset	Training set	Validation set	Test set
100%	60%	20%	20%
3470	2082	694	694

We trained the models on a Windows 10 system with an Intel(R) Core™ i7-8650U processor, 16 GB of random-access memory (RAM), and an NVIDIA GeForce MX150 graphics processing unit (GPU). The models are configured in Python using the Keras version 2.4 API with the TensorFlow version 2.4 backend and CUDA/CuDNN dependencies for GPU acceleration.

#### B. Methods

Our method is segmented into three main parts as shown in Fig. 4, namely preprocessing as seen in the materials section, feature extraction, and classification.



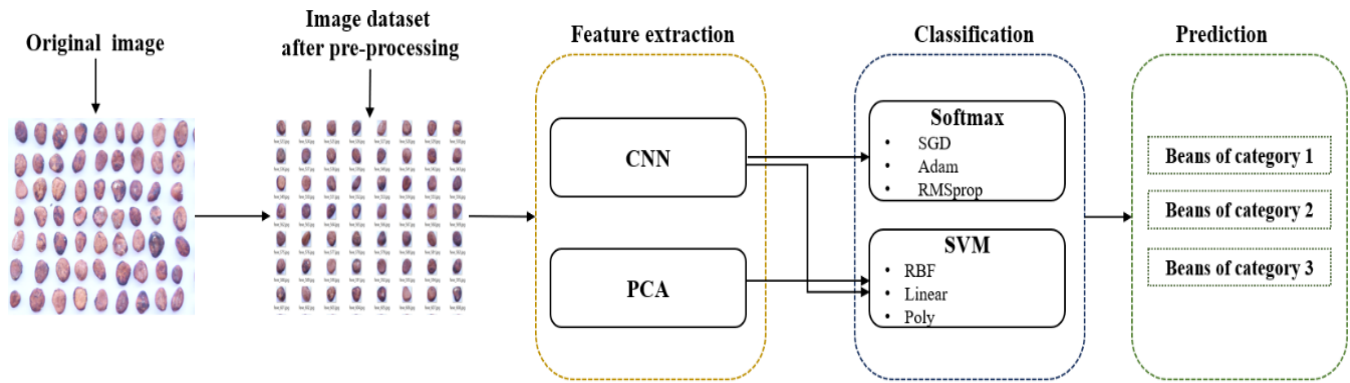


Fig. 4. Architecture of the Model for the Classification of Cocoa Beans.

1) *Feature extraction*: We proposed two feature extractors namely the CNN and the PCA.

a) *CNN*: The CNN we used in our study consists of two convolutional layers with 32 filters each and a 3x3 size kernel with a Relu activation function. The convolutional layer allows for the generation of a particular feature map by applying a filter that examines the whole image. Each of the convolutional layers has a max-pooling layer of 2x2 core, it is a subsampling layer. The subsampling of the pooling layer consists of extracting the most important value of each pattern from the feature map. This layer reduces the parameters and computations in the network, this layer improves the performance of the network and avoids overlearning [15]. Finally, a Flatten layer that flattens the feature map and reduces its size. The Table II gives the description of CNN.

b) *PCA*: Principal component analysis (PCA) is a technique that is applied in applications such as dimensionality reduction, data compression, feature extraction, and data visualization. PCA allows a set of correlated variables X to be transformed into a smaller number y with  $y < X$  of uncorrelated variables called principal components while retaining as much variability of the original data. One of the features of PCA is image compression a technique that reduces the size of an image while retaining as much of the image quality as possible [16].

2) *Classification*: Classification is a task that uses machine learning algorithms that learn to assign a class label to examples in a domain for a given problem. There are many types of classification tasks in machine learning and specialized modeling approaches that can be used for each [17]. In our study, we used the Softmax and SVM classifiers.

a) *Classifier Softmax*: The Softmax classifier is a generalization of the binary form of logistic regression, It has been used in deep learning more precisely in the field of computer vision to classify the vectors obtained after feature extraction [18]. In its operation, the mapping function F is defined such that it takes a set of input data x and maps to output class labels from a simple dot product of the data x and the weight matrix W.

$$F(x_i, W) = W * x_i \quad (1)$$

The Softmax score function gives a probability based on the final score. The sum of the probability of all categories is equal to one [18]. The equation is as follows:

$$f_j(z) = \frac{e^{z_j}}{\sum_k e^{z_k}} \quad (2)$$

The Softmax loss function can be viewed as the entropy of two probabilities, as shown in the following equation:

$$H(p, q) = -\sum_x p(x) \log q(x) \quad (3)$$

We will use optimizers which are algorithms used to minimize the loss function. These functions are:

- SGD which stands for Stochastic Gradient Descent, is a gradient descent optimizer that is used in machine learning and deep learning. Stochastic means a system that is connected or linked with a random probability [19].
- Adam is the extended version of stochastic gradient descent that could be implemented in various deep learning applications such as computer vision and natural language processing [19].
- RMSprop which stands for Root Mean Squared Propagation is an extension of gradient descent and the AdaGrad version of gradient descent that uses a decreasing average of partial gradients to tailor the size of each parameter step. The use of a decreasing moving average allows the algorithm to forget about bad gradients and focus on the best gradients observed during the search progress, overcoming the limitations of AdaGrad [19].

TABLE II. CNN ARCHITECTURE

Layer	Output Shape	Parameter Size
Convolutional 1	(58, 58, 32)	896
Pooling	(29, 29, 32)	0
Convolutional 2	(27, 27, 32)	9248
Pooling	(13, 13, 32)	0
Flattening	5408	0
Total parameter	703 012	
Trainable parameter	703 012	
Non-Trainable parameter	0	



b) *Classifier SVM*: The support vector machine (SVM) is a supervised algorithm used in machine learning. It tries to find a hyperplane that best separates the different data. The SVM is based on statistical approaches, it allows the classification of the data and assigns to each data a specific score as a basis for evaluation. SVM can be used for both regression and classification tasks. However, it is more commonly used for classification objectives [20]. The SVM constructs a hyperplane or a set of hyperplanes in a high or infinite dimensional space, which can be used for classification, regression, or other tasks. Intuitively, a good separation is achieved by the hyper-plane that has the largest distance to the nearest training data points of any class, because in general the larger the margin, the smaller the generalization error of the classifier [21]. The SVM solves the following equation:

$$\min_{\omega, b, \zeta} \frac{1}{2} \omega^T \omega + C \sum_{i=1}^m \zeta_i$$

$$\text{Subject to } y_i(\omega^T \Phi(x_i) + b) \geq 1 - \zeta_i,$$

$$\zeta_i \geq 0, i = 1, \dots, n \quad (4)$$

The loss function is represented by the false predictions of the score function. In SVM, Multiclass SVM Loss is used. The main idea of Multiclass SVM Loss is to determine the scores given by Score Function, requiring the final score to be at least one unit higher than the incorrect score [22]. The loss function is defined by the following equation:

$$L_i = \sum_{j \neq y_i} \max(0, w_j^T x_i - w_{y_i}^T x_i + \Delta) \quad (5)$$

We use the following SVM kernels in our study:

- The linear kernel and its equation is:

$$K(x, y) = x * y \quad (6)$$

- The polynomial kernel and its equation is :

$$K(x, y) = [(x \times y) + 1] d \quad (7)$$

- The RBF (Radial Basis Function) kernel and its equation is:

$$K(x, y) = \exp(-\gamma \|x - y\|^2). \quad (8)$$

Also, we have used the values 1 and 100 for the parameter C which is common to all SVM kernels, it allows us to correct the errors in the classification of the training examples by the simplicity of the decision surface.

### C. Evaluation Metrics

To validate the performance of the pre-trained models in our study we will use the following metrics:

- Accuracy is a performance measure that shows how well the system has classified the data into the correct class.

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN} \quad (9)$$

- Precision is the ratio of correctly classified positive images to the total number of true positive images.

$$\text{Precision} = \frac{TP}{TP+FP} \quad (10)$$

- Recall is the ability of a classifier to determine actual positive outcomes

$$\text{Recall} = \frac{TP}{TP+FN} \quad (11)$$

- The F1 score is the weighted average of precision and recall

$$F1 \text{ score} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad (12)$$

- The Matthews correlation coefficient (MCC) is used in machine learning as a measure of the quality of classifications

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad (13)$$

- The mean square error of an estimator measures the average of the squared errors, i.e. the mean square difference between the estimated values and the true value.

$$MSE = \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 \quad (14)$$

With TP: True positive, TN: True negative, FP: False positive, and FN: False negative.

## IV. RESULTS

### A. CNN with Softmax

In the case of the CNN with Softmax, we can record the results presented in the Table III which takes into account the different optimizers mentioned above.

TABLE III. THE GENERAL PERFORMANCE OF THE SOFTMAX CLASSIFIER

	Accuracy	Loss	Precision	F1 score	Recall	MCC	MSE
SGD	92,95	17,93	93,37	92,95	92,95	89,2	3,41
Adam	<b>95,64</b>	<b>17,18</b>	<b>95,65</b>	<b>95,64</b>	<b>95,63</b>	<b>93,18</b>	<b>2,37</b>
RMSprop	94,63	36,99	94,83	94,62	94,63	91,73	3,18

The results presented in Table III show that the Adam optimizer obtains the best performance with a score of 95.64%, followed by the RMSprop and finally the SGD. We see in this experiment that the softmax classifier using the Adam optimizer is more optimal.

Fig. 5 presents the confusion matrix of each softmax experiment case; also, a histogram of the metrics has been created.

### B. SVM

In the case of SVM we can record the results presented by the Table IV which takes into account the different kernels and parameters.

The results presented in Table IV show that the SVM with the RBF kernel; and the C parameter at 100 obtained the best performance with a score of 97.32%.

We now present the confusion matrices and histograms of the metrics in Fig. 6.

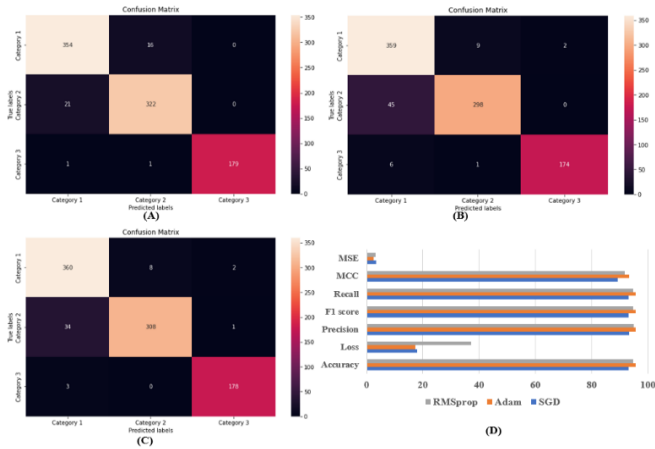


Fig. 5. (A) Confusion Matrix of Adam ; (B) Confusion Matrix of SGD ; (C) Confusion Matrix of RMSprop ; (D) Graphical Representation of Optimizer Metrics.

TABLE IV. GENERAL PERFORMANCE OF THE SVM CLASSIFIER

	Accuracy	Precision	F1 score	Recall	MCC	MSE
<i>SVM (rbf ; C=1)</i>	93,96	94,08	93,94	93,95	90,65	6,37
<i>SVM (rbf ; C=100)</i>	<b>97,32</b>	<b>97,33</b>	<b>97,31</b>	<b>97,31</b>	<b>95,82</b>	<b>3,35</b>
<i>SVM (linear ; C=1)</i>	93,4	93,55	93,38	93,4	89,79	7,94
<i>SVM (linear ; C=100)</i>	91,39	91,52	91,36	91,38	86,64	9,95
<i>SVM (poly ; C=1)</i>	95,75	95,82	95,74	95,74	93,41	5,25
<i>SVM (poly ; C=100)</i>	95,97	96,02	95,97	95,97	93,73	5,03

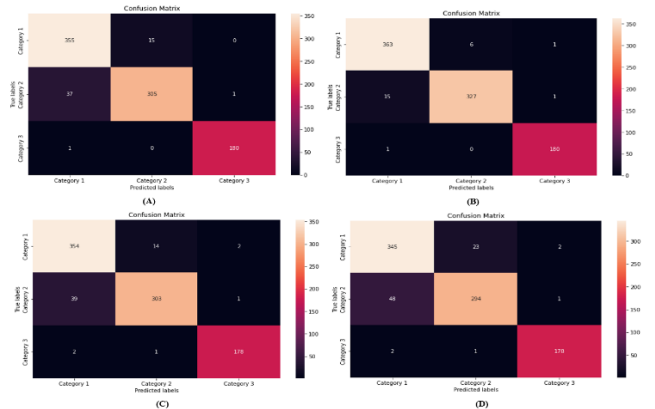


Fig. 6. (A) Confusion Matrix of SVM (RBF ; C=1) ; (B) Confusion Matrix of SVM (RBF ; C=100) ; (C) Confusion Matrix of SVM (Linear ; C=1) ; (D) Confusion Matrix of SVM (Linear ; C=100) ; (E) Confusion Matrix of SVM (Poly ; C=1) ; (F) Confusion Matrix of SVM (Poly ; C=100) ; (G) Graphical Representation of Optimizer Metrics.

### C. CNN with SVM

For the hybrid CNN-SVM method, we can record the results presented in Table V which takes into account the different kernels and SVM parameters.

TABLE V. GENERAL PERFORMANCE OF THE CNN WITH SVM

	Accuracy	Precision	F1 score	Recall	MCC	MSE
<i>CNN - SVM (rbf ; C=1)</i>	95,08	95,29	95,06	95,07	92,44	4,92
<i>CNN - SVM (rbf ; C=100)</i>	<b>98,32</b>	<b>98,34</b>	<b>98,32</b>	<b>98,32</b>	<b>97,39</b>	<b>1,67</b>
<i>CNN - SVM (linear ; C=1)</i>	95,86	95,95	95,85	95,86	93,59	4,47
<i>CNN - SVM (linear ; C=100)</i>	95,86	95,95	95,85	95,86	93,59	4,47
<i>CNN - SVM (poly ; C=1)</i>	95,41	95,61	95,4	95,41	92,96	5,25
<i>CNN - SVM (poly ; C=100)</i>	98,10	98,11	98,09	98,09	97,04	1,90

The results presented in Table V show that the CNN-SVM with the RBF kernel; and the C parameter at 100 obtained the best performance with a score of 98.32%.

We now present the confusion matrices and histograms of the metrics in Fig. 7.

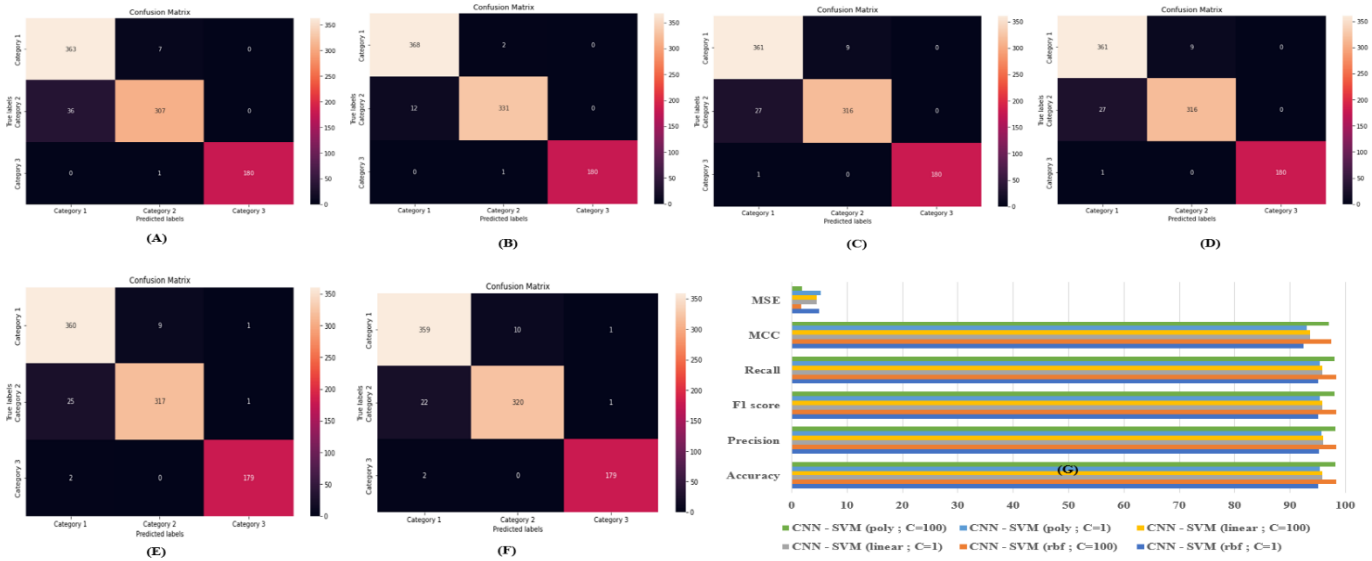


Fig. 7. (A) Confusion Matrix of CNN - SVM (rbf ; C=1); (B) Confusion Matrix of CNN - SVM (rbf ; C=100); (C) Confusion Matrix of CNN - SVM (Linear ; C=1); (D) Confusion Matrix of CNN - SVM (Linear ; C=100); (E) Confusion Matrix of CNN - SVM (Poly ; C=1).

D. PCA with SVM

For the PCA-SVM method, we can record the results presented in Table VI which takes into account the different kernels and parameters of the SVM.

The results presented in Table VI show that the SVM with the RBF kernel; and the C parameter at 100 obtained the best performance with a score of 97.65%.

We now present the confusion matrices and histograms of the metrics in Fig. 8.

E. Comparison of the Results of the Different Methods

We will compare the results of the different methods used in our study, namely: CNN, CNN-SVM, SVM, and PCA-SVM. The results will be represented in Table VII.

Table VII compares the best results obtained in our different experiments. It appears that the hybrid CNN-SVM method obtained the best score followed by the PCA-SVM. Fig. 9 shows the histogram of the comparison

TABLE VI. GENERAL PERFORMANCE OF THE PCA WITH SVM

	Accuracy	Precision	F1 score	Recall	MCC	MSE
PCA - SVM (rbf ; C=1)	95,41	95,49	95,4	95,41	92,89	5,21
PCA - SVM (RBF; C=100)	97,65	97,67	97,64	97,65	96,34	3,02
PCA - SVM (linear; C=1)	93,40	93,55	93,38	93,4	89,79	7,94
PCA - SVM (linear; C=100)	91,39	91,52	91,36	91,38	86,64	9,95
PCA - SVM (poly; C=1)	92,73	93,13	92,74	92,72	88,81	10,96
PCA - SVM (poly; C=100)	97,20	97,2	97,19	97,2	95,64	3,13

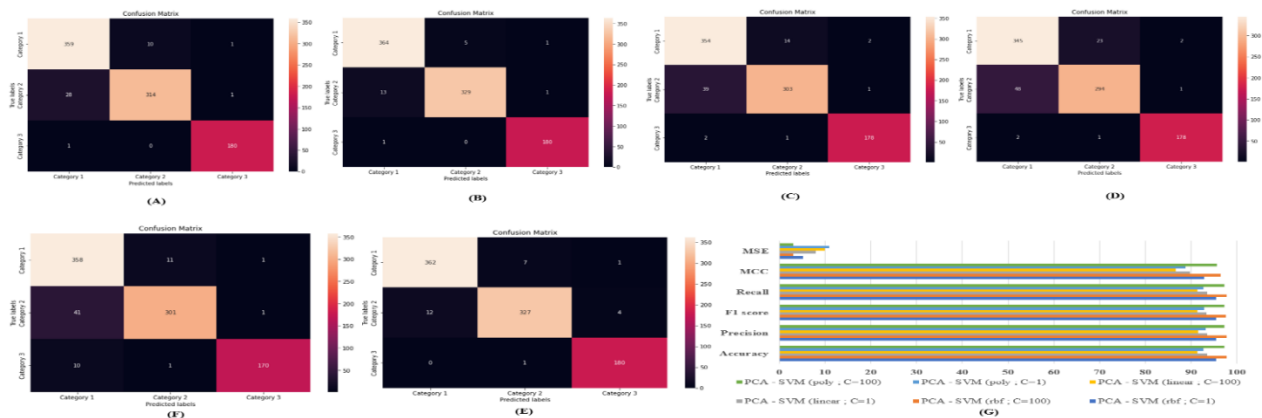


Fig. 8. (A) Confusion Matrix of PCA - SVM (rbf ; C=1); (B) Confusion Matrix of PCA - SVM (rbf ; C=100); (C) Confusion Matrix of PCA - SVM (Linear ; C=1); (D) Confusion Matrix of PCA - SVM (Linear ; C=100); (E) Confusion Matrix of PCA - SVM (Poly ; C=1); (F) Confusion Matrix of PCA - SVM (Poly ; C=100); (G) Graphical Representation of Optimizer Metrics.

TABLE VII. COMPARATIVE TABLE OF THE BEST RESULTS OF OUR EXPERIMENTS

Models	Accuracy
CNN - SVM (RBF; C=100)	98,32
PCA - SVM (RBF; C=100)	97,65
SVM (RBF; C=100)	97,32
Softmax with Adam optimizer	95,64

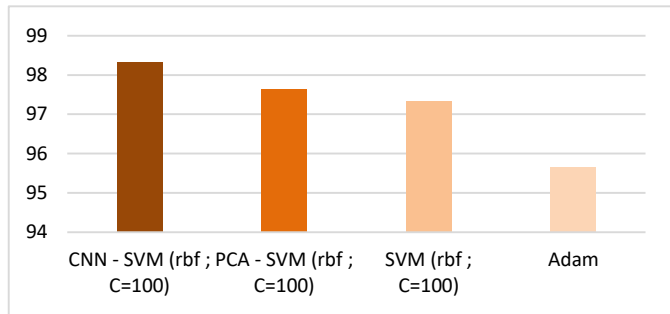


Fig. 9. Histogram of Best Scores of Experiments.

#### F. State-of-the-art Comparison

The results of our experiments have given better results than the state of the art and Table VIII presents the results. These results are also represented by the histogram as shown in Fig. 10.

TABLE VIII. COMPARISON OF THE RESULTS OF OUR EXPERIMENTS WITH THE STATE-OF-THE-ART

Models	Accuracy
Oliviera et al. [9]	92
CNN - SVM (rbf ; C=100)	98,32
PCA - SVM (rbf ; C=100)	97,65
SVM (rbf ; C=100)	97,32
Softmax with Adam optimizer	95,64

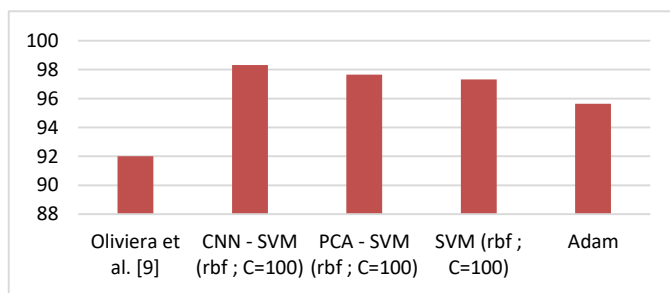


Fig. 10. Histogram of the Results of our Experiments with the State-of-the-Art.

#### V. DISCUSSION

The analysis of digital images of a cocoa bean using CNN and PCA-based feature extractors were used to then perform the classification of cocoa beans from softmax and SVM classifier. The work resulted in the following:

First, we used the CNN coupled with the softmax classifier using several optimizers in this case Adam, RMSprop, and

SGD. The Adam optimizer obtains the best performance with a score of 95.64%, followed by RMSprop and SGD. These results are presented in Table III. In the second step, we used the SVM classifier using several kernels such as rbf, linear, and poly with the parameters C with a value of 1 and 100. We thus obtained a score of 97.32% which represents the best accuracy coming from the rbf kernel with C = 100, these results are presented in Table IV. In a third step, we used the hybrid CNN-SVM method with the same parameters used by the SVM, again we have a score of 98.32% achieved with the rbf kernel and C=100, these results are presented in Table V. Finally in a fourth time we have the PCA-SVM method which also uses the same parameters of the SVM and obtained a score 97.65% with the kernel rbf and C=100, these results are presented in Table VI. At the end of our work, we realize that the hybrid CNN-SVM method obtained the best accuracy of all the methods used as shown in Table VII of the comparative study, and also of the methods of previous studies in the literature review as shown in Table VIII.

#### VI. CONCLUSION

Cocoa production is an area of research that needs the use of automated methods for product quality assessment. The results obtained showed that feature extractions based on CNN coupled with an SVM classifier are promising systems to classify cocoa beans according to quality. Also, we used principal component analysis to reduce the size of our data while designing the main features and this allowed us to have a satisfactory result, which results in showing that we can minimize the computational time of the classifiers proceeding to a reduction of the dimensions. The result of the hybrid CNN-SVM method obtained the best score of all the methods used including the one in the literature. We achieved our goal because the hybrid method gives us a better score. In future work, we will be able to use texture extraction methods and pre-trained CNN sets for more accuracy. Also, a study of the quality of cocoa beans based on the approach of the maturity of cocoa pods will initially distinguish the best pods. The harvest of unripe pods gives beans of poor quality, while a pod too ripe has beans that begin to germinate or alter inside the pod. The classification of cocoa beans according to their shape or morphological parameters will also help to obtain quality beans.

#### ACKNOWLEDGMENT

Our gratitude goes to the village community of YAKASSE, located in the town of GRAND-BASSAM in CÔTE D'IVOIRE, who kindly put their plantations at our disposal for our studies.

#### REFERENCES

- [1] « Fèves de cacao : production mondiale par pays 2022 », Statista. <https://fr.statista.com/statistiques/565101/production-mondiale-feves-cacao-volume-par-pays/> (consulté le 30 juillet 2022).
- [2] Nair, K. P. (2021). Cocoa (Theobroma cacao L.). In *Tree Crops* (pp. 153-213). Springer, Cham.
- [3] Urbańska, B., & Kowalska, J. (2019). Comparison of the total polyphenol content and antioxidant activity of chocolate obtained from roasted and unroasted cocoa beans from different regions of the world. *Antioxidants*, 8(8), 283.

- [4] L. Samagaci, H. Ouattara, S. Niamké, et M. Lemaire, « Pichia kudrazevii and Candida nitrativorans are the most well-adapted and relevant yeast species fermenting cocoa in Agneby-Tiassa, a local Ivorian cocoa producing region », *Food Res. Int.*, vol. 89, p. 773-780, nov. 2016, doi: 10.1016/j.foodres.2016.10.007.
- [5] Engeseth, N. J., & Pangan, M. F. A. (2018). Current context on chocolate flavor development—A review. *Current opinion in food science*, 21, 84-91.
- [6] De Vuyst, L., & Leroy, F. (2020). Functional role of yeasts, lactic acid bacteria and acetic acid bacteria in cocoa fermentation processes. *FEMS Microbiology Reviews*, 44(4), 432-453.
- [7] S. Sood et H. Singh, « Computer Vision and Machine Learning based approaches for Food Security: A Review », *Multimed. Tools Appl.*, vol. 80, no 18, p. 27973-27999, juill. 2021, doi: 10.1007/s11042-021-11036-2.
- [8] Ghorbanzadeh, O., Blaschke, T., Gholamnia, K., Meena, S. R., Tiede, D., & Aryal, J. (2019). Evaluation of different machine learning methods and deep-learning convolutional neural networks for landslide detection. *Remote Sensing*, 11(2), 196.
- [9] M. M. Oliveira, B. V. Cerqueira, S. Barbon, et D. F. Barbin, « Classification of fermented cocoa beans (cut test) using computer vision », *J. Food Compos. Anal.*, vol. 97, p. 103771, avr. 2021, doi: 10.1016/j.jfca.2020.103771.
- [10] B. Khagi, C. G. Lee, et G.-R. Kwon, « Alzheimer's disease Classification from Brain MRI based on transfer learning from CNN », in *2018 11th Biomedical Engineering International Conference (BMEiCON)*, Chiang Mai, nov. 2018, p. 1-4. doi: 10.1109/BMEiCON.2018.8609974.
- [11] A. P. A. C. Barbon, S. Barbon, R. G. Mantovani, E. M. Fuzyi, L. M. Peres, et A. M. Bridi, « Storage time prediction of pork by Computational Intelligence », *Comput. Electron. Agric.*, vol. 127, p. 368-375, sept. 2016, doi: 10.1016/j.compag.2016.06.028.
- [12] Muthulakshmi. A et P. N. Renjith, « Classification of Durian Fruits based on Ripening with Machine Learning Techniques », in *2020 3rd International Conference on Intelligent Sustainable Systems (ICISS)*, Thoothukudi, India, déc. 2020, p. 542-547. doi: 10.1109/ICISS49785.2020.9316006.
- [13] B. Harel, Y. Parmet, et Y. Edan, « Maturity classification of sweet peppers using image datasets acquired in different times », *Comput. Ind.*, vol. 121, p. 103274, oct. 2020, doi: 10.1016/j.compind.2020.103274.
- [14] M. Elleuch, R. Maalej, et M. Kherallah, « A New Design Based-SVM of the CNN Classifier Architecture with Dropout for Offline Arabic Handwritten Recognition », *Procedia Comput. Sci.*, vol. 80, p. 1712-1723, 2016, doi: 10.1016/j.procs.2016.05.512.
- [15] O. Alharbi, « A Deep Learning Approach Combining CNN and Bi-LSTM with SVM Classifier for Arabic Sentiment Analysis », *Int. J. Adv. Comput. Sci. Appl.*, vol. 12, no 6, 2021, doi: 10.14569/IJACSA.2021.0120618.
- [16] M. Usman, S. Ahmed, J. Ferzund, A. Mehmood, et A. Rehman, « Using PCA and Factor Analysis for Dimensionality Reduction of Bio-informatics Data », *Int. J. Adv. Comput. Sci. Appl.*, vol. 8, no 5, 2017, doi: 10.14569/IJACSA.2017.080551.
- [17] Kowsari, Jafari Meimandi, Heidarysafa, Mendu, Barnes, et Brown, « Text Classification Algorithms: A Survey », *Information*, vol. 10, no 4, p. 150, avr. 2019, doi: 10.3390/info10040150.
- [18] X. Qi, T. Wang, et J. Liu, « Comparison of Support Vector Machine and Softmax Classifiers in Computer Vision », in *2017 Second International Conference on Mechanical, Control and Computer Engineering (ICMCCE)*, Harbin, déc. 2017, p. 151-155. doi: 10.1109/ICMCCE.2017.49.
- [19] A. Chadha, A. Abdullah, et L. Angeline, « A Comparative Performance of Optimizers and Tuning of Neural Networks for Spoof Detection Framework », *Int. J. Adv. Comput. Sci. Appl.*, vol. 13, no 4, 2022, doi: 10.14569/IJACSA.2022.01304102.
- [20] M. Ahmad, S. Aftab, M. Salman, N. Hameed, I. Ali, et Z. Nawaz, « SVM Optimization for Sentiment Analysis », *Int. J. Adv. Comput. Sci. Appl.*, vol. 9, no 4, 2018, doi: 10.14569/IJACSA.2018.090455.
- [21] Müller, K. R., Mika, S., Tsuda, K., & Schölkopf, K. (2018). An introduction to kernel-based learning algorithms. In *Handbook of Neural Network Signal Processing* (pp. 4-1). CRC Press.
- [22] S. Ghosh, A. Dasgupta and A. Swetapadma, "A Study on Support Vector Machine based Linear and Non-Linear Pattern Classification," *2019 International Conference on Intelligent Sustainable Systems (ICISS)*, 2019, pp. 24-28, doi: 10.1109/ISS1.2019.8908018.

# SQrum: An Improved Method of Scrum

## Proposed Metamodel for SQrum Method

Najihi Soukaina, Merzouk Soukaina, Marzak Abdelaziz

Department of Mathematics and Computer Sciences, Hassan II University- Casablanca  
Faculty of Sciences Ben M'sik, Casablanca, Morocco

**Abstract**—Software systems are having a major impact on many aspects of personal and professional life. Safety-critical applications, such as production line controls, automotive operations, and process industry controls, rely significantly on software systems. In these applications, software failure may result in bodily damage or death. The proper operation of software is essential to the safety and well-being of individuals and businesses. Therefore, software quality assurance is of paramount relevance in the software business today. In recent years, Agile Project Management and particularly Scrum, have gained popularity as a method of dealing with "vuca" business environments, which are characterized by rising Volatility, Uncertainty, Complexity, and Ambiguity. This paper contributes to the software development body of knowledge by proposing a metamodel of Scrum quality assurance, named SQrum ('SQ' of Software Quality and 'rum' of Scrum). Our objective is to make Scrum more efficient and reliable and to assist enterprises in undertaking quality assurance activities while considering agile practices and values.

**Keywords**—Agile project management; IT; OMG; meta-object facility; MOF; metamodel; scrum; SQrum; quality assurance; QA; quality management; QM; software development project

### I. INTRODUCTION

The capacity to successfully execute Information Technology (IT) projects has become a crucial and strategic need for modern organizations. When expenses for field updates, recalls, repairs, downtime, etc. are included, the release of a product with problems may be extremely costly. Damage to a company's reputation is less measurable but equally significant. In addition, a failed software project might damage the competitive position of an organization. Value and quality delivery are essential variables for determining the success of a software development project; they are crucial assets for any firm seeking to remain competitive in the marketplace. Despite extensive efforts to find methods for maintaining software quality, software projects continue to fail [1].

There is a growing demand for the deployment of software development methods that are both flexible enough to keep up with the rapid rate of change and the competitive market and rigorous enough to prevent defects and assure product quality. However, humans are fallible. Even with the most advanced and conscientious design processes, erroneous outcomes cannot be avoided beforehand. As a result, software products, like the outcomes of any engineering effort, must be validated against their requirements throughout their development. Agile software development has arisen as an alternative to planning

and managing complicated projects by offering methods to accommodate frequent project changes.

Agile is a method characterized by continuous iterations and testing throughout a product's Software Development Life Cycle. Scrum is the most popular agile approach [2]. Scrum is a lightweight, agile framework that provides processes for managing and controlling the software and product development process. Although Scrum offers a number of benefits, such as incremental deliverables at the end of each iteration, stakeholders and product owners can modify requirements throughout the process, and Scrum can swiftly adapt to these modifications. Process and product quality remain Scrum's principal problems.

One of the most significant contrasts between agile and traditional development is the agile "whole-team" approach [3], in which quality is the responsibility of the entire team and quality assurance is incorporated into the process itself, without an explicit Quality Manager (QM) role. Quality Assurance (QA) activities are integrated into the team's day-to-day operations in order to improve product quality through a smooth process. Large organizations frequently face the issue of combining the Agile requirements of adaptability, transparency, and collaboration while also assuring product quality and adhering to required QA processes. With these changing issues in this area, businesses struggle to identify the best method to integrate QA into Agile environments, particularly Scrum.

Within an Agile team, each team member is responsible for testing and product quality and participates in test-related activities. Each member of the team may see quality from a unique perspective and mindset. All of them are acceptable contributions to quality, but Scrum projects are still facing quality challenges due to the lack of defined rules in Scrum. This necessitates an explicit QM position to incorporate non-functional requirements (NFR), assure process improvement, and provide shared ownership of quality. Just as the Product Owner is responsible for maximizing the value, the QM is responsible for increasing the quality.

Every member of an agile team is a tester, but a QM is more than just a tester. A QM on the Agile team is able to give an overview perspective on all team contributions in order to establish the product quality strategy. Instead of criticizing, QM provides proactive ways to improve productivity, promote software quality within the team, and give software testing and quality coaching.



Quality assurance is like a ship; everyone participates in making it move forward, but only one person can steer it in the right direction. A QM is the one who gives the "course" to follow for the whole team to reach a satisfactory quality level. He is the one who carries the global vision of the product and process quality.

The aim of this paper is to propose a new agile approach called "SQrum", whose central concept is the addition of a new quality role and all of the artifacts it will require. We will build a metamodel using the OMG Meta-Object Facility (MOF) that provides an abstract view of the new agile development process. The Meta Object Facility (MOF [4]) standard was created by the Object Management Group (OMG) to facilitate the development of modeling formalisms in the form of metamodels. This consists of a set of meta-classes connected by meta-association [11]. The rest of this paper is organized as follows: Section II details the related work of this study. Section III provides an overview of Scrum. Section IV describes the disadvantages of Scrum. SQrum is presented in Section V. Finally, Section VI concludes the study.

## II. RELATED WORK

Hanssen et al. [5] found that Quality Assurance processes in Scrum are becoming inadequate. Consequently, they suggested a new method called SafeScrum, which is a Scrum variant with some supplementary XP methods that can be used to develop safety-critical software and IEC 61508 certified software. They evaluated the standard, sought out an impartial evaluator, and collaborated with the Scrum team to identify the required additional tasks to be included in the Safe Scrum process for an internal quality assurance component. They did not cover all aspects of quality in a scrum project with the proposed role.

Jeon et al. [6] explain that Scrum does not place enough focus on non-functional requirements, whereas the key success factor of software projects is not only the satisfaction of functionalities, but also of quality attributes; therefore, they propose the ACRUM method for the analysis and incorporation of quality attributes into software projects. Jan Bosch concurs with Jeon et al. and suggests an approach that explicitly takes nonfunctional criteria into account during design.

Timperi [7] explains the weakness of scrum is the lack of concrete guidance and instructions about quality assurance activities and that the focus has been on the development activities while quality assurance practices of different agile methodologies have received less attention and an overall picture is missing. The author recommends combining quality assurance practices of different methodologies, like Scrum and XP, in order to get good enough software delivered to the customer.

Aamir et al. [8] assert that due to the rapid delivery process of sprints, quality is not considered in the scrum framework, and the majority of Quality Control (QC) operations are overlooked. To address this issue, the authors offer an enhanced scrum model for implementing QC activities and evaluating the product's quality. They also provide a new

concept of "test backlog" for documenting test cases within the scrum.

Bajnaid et al. [9] addressed the limitations of agile practices that do not include quality assurance in their process to guarantee that the quality assurance procedure has been followed and quality assurance criteria have been satisfied. To overcome the drawbacks, they suggested a process-driven e-learning system that senses developers' activities and guides them through necessary software quality assurance methods during software development.

## III. SCRUM OVERVIEW

Scrum is a management process that was initially mentioned in 1986 by Takeuchi and Nonaka in their paper "The New New Product Development Game," in which they describe a flexible, fast, and self-organizing product development process. Sutherland and Schwaber [10] used these discoveries and the word "Scrum" to create the currently known framework, which was initially introduced in 1995.

Scrum is a lightweight development methodology that allows IT organizations to handle complicated adaptive challenges while delivering solutions of the highest possible quality. Scrum was developed based on the premise that software development is too complicated and unpredictable to be meticulously planned at the start of a project [11]. Therefore, a progressive, iterative method is used to maximize predictability and limit risk. Given that change cannot be avoided, it must be managed. Scrum handles change by developing software in iterations as opposed to a one-shot method.

Scrum is based around three roles (Product Owner, Scrum Master, and Development Team), four meetings (Sprint Planning, Daily Scrum, Sprint Review and Sprint Retrospective) and three artifacts (Product Backlog, Sprint Backlog, Product Increment) [10] (see Fig. 1.)

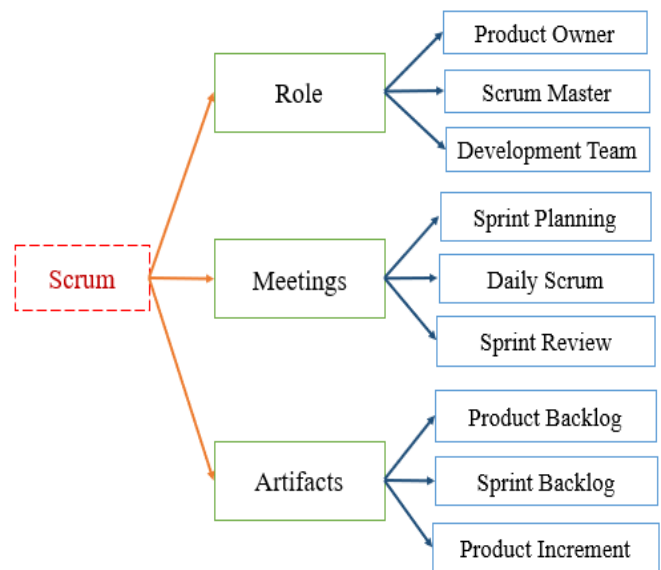


Fig. 1. Roles, Artifacts, and Activities in Scrum [12].

Each project has a Product Backlog, which is a list of a product's requirements ordered by business value. It is a constantly evolving document based on changing requirements, changing problem understanding, and changing contexts. Each Product Backlog Item (PBI) is estimated using an abstract effort measure based on "Story Points" and has a set of acceptance criteria.

Scrum teams are self-organized, multi-skilled, and capable of producing products iteratively and incrementally, thus increasing opportunities for ongoing feedback. A Scrum team consists of a product owner, who acts on behalf of the client and is charged with maximizing the value of the developed product, and a development team, which is responsible for creating the product. The development team is made up of developers and a Scrum master. The Scrum master is a facilitator who ensures that the development team is supplied with an appropriate environment to finish the project effectively, removes impediments for the team, and guarantees adherence to Scrum practices.

The Scrum development process is carried out by cross-functional teams of individuals with diverse skill sets[10]. The teams often possess a variety of specializations, including programming, testing, analysis, database administration, user experience, and infrastructure. All of these skills are required to provide the product, and Agile projects employ a whole-team approach to execute it. Advantages of employing a whole-team approach include the fact that quality is everyone's responsibility. Scrum focuses on developing high-quality software within a timeframe that optimizes its business value. This is everyone's responsibility, not just the testers. Every member of a scrum team is a tester. Tests, from the unit level on up, drive the code, teach the team how the program should function, and indicate when a task or story is "done." A Scrum team must have all the competences necessary to generate high-quality code that provides the organization's requested features. This means that the team is responsible for all testing activities, including test automation and manual exploratory testing. It also implies that the entire team continuously considers testability while creating code.

Scrum divides a project into iterations known as "Sprints". A Sprint is a time-boxed, often 30-day iteration in which the

Scrum team adds new features to the product. The sprint begins with a "Sprint Planning Meeting" where the team picks from the product backlog the items to be handled in the sprint and plans the work to be performed. The team will estimate the selected items based on their velocity (e.g., the number of "story points" they can execute within a predetermined time limit). The result of planning is turned into an objective known as the "Sprint Goal" The Scrum Team then has an internal meeting and utilizes the Sprint Goal to generate a list of the necessary requirements to achieve the target. These requirements are decomposed into "tasks" that become entries in the "Sprint Backlog." The success of the sprint is based on the achievement of the sprint goal [2].

During the sprint, the team holds daily 15-minute stand-up meetings called "Daily Scrum" with the purpose of assessing progress and maximizing the chance that the development team will accomplish the sprint goal. Each team member responds to three questions [13]:

- What progress has been made since the last meeting?
- What will be accomplished by the next meeting?
- What impediments stand in the way?

The Sprint yields a deliverable product increment as its final output. During a 4-hour Sprint Review, the Scrum Team inspects the product increment, evaluates what they were able to accomplish during the sprint, and modifies the product backlog as needed prior to the next Sprint Planning meeting [17]. The "Product Owner" will approve the needs for the live system based on the requirements and their preset acceptance criteria.

The last meeting of a "sprint" is the "retrospective," at which time the team examines and comments on itself and the project in terms of people, relationships, processes, and tools. As a consequence, an improvement strategy may be developed.

Merzouk et al. [12] proposed, in their paper titled "Towards a New Metamodel Approach of Scrum, XP, and Ignite Methods," a metamodel of Scrum (see Fig. 2).

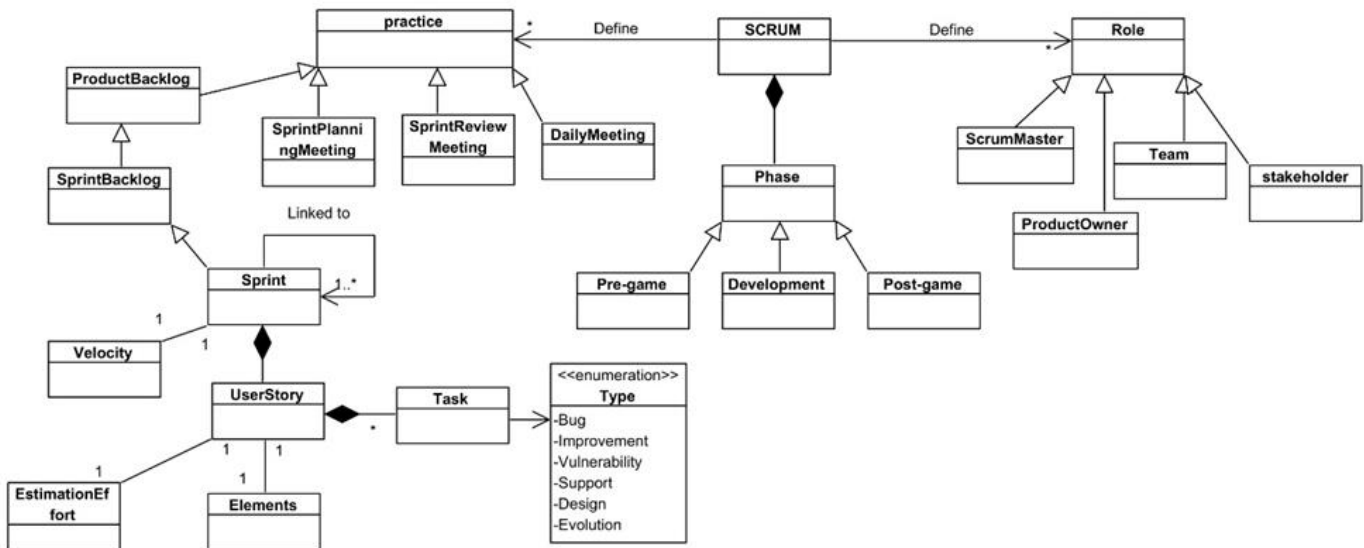


Fig. 2. Proposed Metamodel for Scrum Method [12]

#### IV. CRITICAL ANALYSIS OF SCRUM

Scrum is the most popular and effective Agile methodology due to its many advantages, including quick delivery and flexibility to change. It emphasizes customer satisfaction, continuous feedback, and process transparency. In spite of its many advantages, it has three disadvantages. First, Scrum backlogs focus only on functional requirements (FR) and tend to neglect non-functional requirements (aka Quality Attributes; see Fig. 3) [6] [14]. Second, the majority of quality assurance activities are skipped in scrum due to the sprint's short period and the lack of a dedicated quality management role [8]. Finally, in Scrum, the requirements are typically managed by a person with a business-oriented profile. Thus, the focus is on the development activities that produce business value, while

quality assurance practices receive less attention and an overall picture is missing [7].

This study proposes the SQRum as the solution to the aforementioned problems. SQRum is built on the traditional scrum to present a more effective method.

Our new approach will improve classic Scrum in three ways.

- The quality attributes will be considered and included in the product backlog.
- Quality assurance activities will be effectively handled.
- The tester's responsibilities will be precisely outlined.

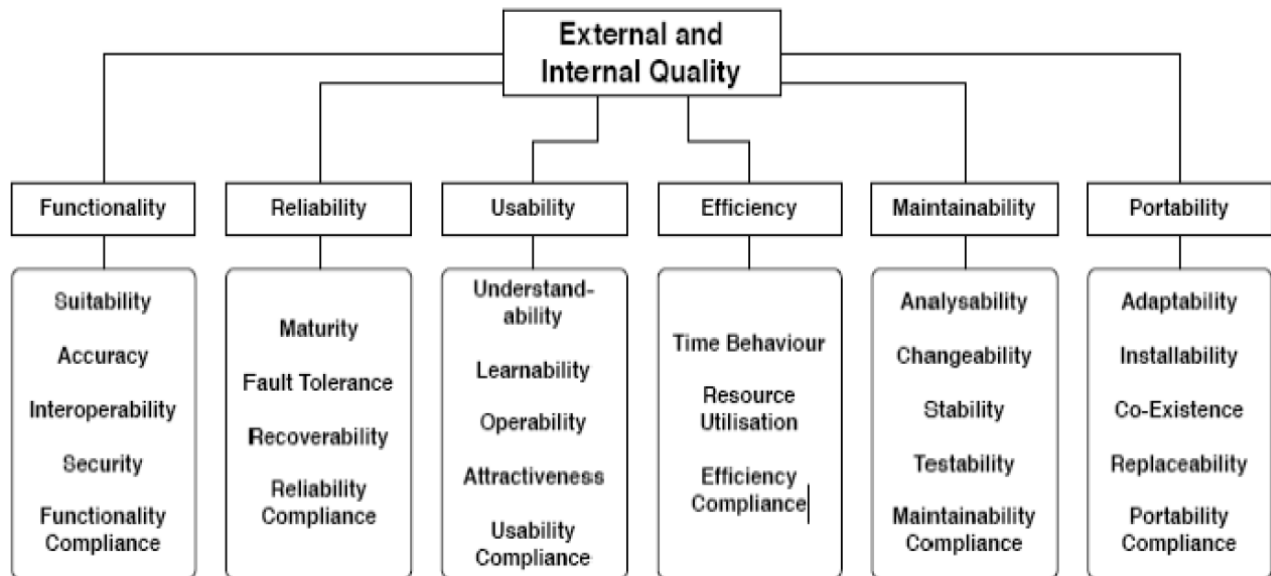


Fig. 3. ISO 9126 Model for Quality Attribute [15].

## V. PROPOSED METHOD

### A. SQRum Quality Artifacts

SQRum intends to incorporate the DoD as an artifact into its new model and proposes four new artifacts: QA strategy, test plan, defect backlog, and test library.

1) *Definition of done*: DoD is a significant part of the test plan. It is used by the QM as a check list of items, each one used to validate a story or a PBI for completeness. Unlike the acceptance criteria, the DoD is applicable to all items in the Product Backlog, not just a single user story. It is applied to the product increment as a whole [16]. The QM collaborates with the PO and the dev team to identify all the conditions that make an increment shippable or not at the end of the sprint. This proposal's DoD should emphasize quality attributes.

2) *QA strategy*: A QA strategy is a long-term plan of action, the key word being "long-term" about the overall test approach to projects. It defines the project's testing guidelines on how to test the target system.

Using the ship analogy once more, the QA strategy symbolizes the "course" to follow in order to reach the final destination.

QA strategy is a static high-level document that can be used as a reference that doesn't change much over time and needs to be updated only if processes change. It generally includes decisions made in terms of sprint timelines, test types required, infrastructure such as test management tools, defects management tool, test environments, test monitoring and reporting.

3) *Test plan*: A test plan is a concise and lightweight document used to organize the test activities. Each sprint has its own test plan, which is a living document that changes and evolves based on sprint requirements. A test plan outlines the scope, approach, resources, and schedule of planned testing activities. It identifies, among other things, the items and features to be tested, the assignment of tasks, the DoD and the test types to be performed, the test environment, the test design techniques, test data requirements and test measurement techniques to be used, the risk and dependencies assessment carried out, automation tests programmed, and time budget allocated. Continuing with the ship metaphor, the sprint is the trip, and the test plan is all that is required to ensure its success.

An Agile Test Plan is a crucial document since it collects all the answers to test-related questions in one place.

4) *Defect backlog*: A defect backlog is an ordered list of all the known defects in the project that haven't been fixed yet.

Defects may be functional, describing misbehavior or technical related to quality attributes such as performance, security, or other. The remaining bugs can be calculated by subtracting the fixed bugs from the total bugs:

$$\text{Remaining bugs} = \text{number of total bugs} - \text{fixed bugs}$$

There are three types of defects that we can find in the defect backlog:

- **Defects within the current iteration**: These are defects that can't be fixed immediately and do not impact the increment date of release. Ideally, defects should be corrected as soon as they are discovered, before they become massive, tangled defects.
- **From the Legacy System**: These are inherited defects of the old system that have remained hidden until now. When found, they are logged to the defect backlog and the QM with the team can choose to fix them or not. If so, they will be prioritized as part of the product backlog.
- **Found in Production**: These are bugs found by the customer in production. Depending on their severity, these bugs may be fixed immediately, at the time of the next release, or they'll be estimated, prioritized, and put in your product backlog.

5) *Test library*: Scrum uses an iterative approach for product development; the same approach can also be applied to testing. As a product is produced, the test library expands progressively. It includes test cases, test scenarios, and test campaigns. User stories are the basis for the creation of test cases. The relationship between test cases and user stories in the sprint backlog is one-to-many, as a single user story may have numerous test cases. The test cases are similar to puzzle pieces that compose the test scenario (see Fig. 4), and test scenarios are the main components of test campaigns. Each campaign is designed to evaluate an increment. Utilizing the library enables testers to save time since for each new script, there are reusable components that can be used to develop a new test campaign. Also, all they have to do is make test cases and scenarios for the new features. For example, a first sprint is conducted to develop a command-launching feature. The test campaign is then utilized to check that this increment functions properly. Sprint 2: This sprint's objective is to develop functionality that allows for the management of billing for placed orders. Testers won't have to start from scratch to test this new feature. Instead, they can use the first tests they ran to launch the order and finish this test campaign with tests that make sure billing is handled correctly.

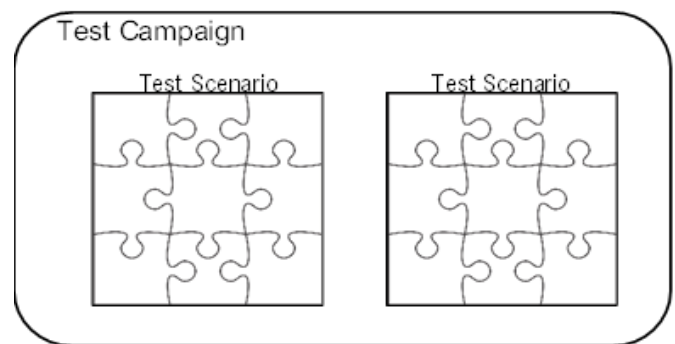


Fig. 4. The Components of a Test Campaign.

## B. SQRum Events

1) *Refactoring iteration*: it is a dedicated sprint for continuously enhancing the design of an existing solution without modifying its core behavior. Agile teams incrementally maintain and enhance their code from Sprint to Sprint. If code is not refactored, the resulting product will be of poor quality, with unhealthy dependencies between building blocks and improper allocation of component responsibilities [17].

2) *Product backlog grooming*: is a meeting that helps the Product Owner maintain his Product Backlog. He can ask for the aid of the team in creating and refining Product Backlog Items, estimating the amount of work required to complete a PBI, and prioritizing PBIs in the Product Backlog to ensure that the Product Backlog is ready according to the Definition of Ready [18].

## C. Quality Manager Responsibilities

1) *Manage quality attributes*: Quality Manager is responsible for NFR. Also known as quality attributes (see Fig. 3), QM works with the scrum team to analyze the quality attributes based on the functional requirements. With the agreement of the product owner, he completes the product backlog by mapping out the functional requirements and the quality attributes. QM can use three ways to elicit NFR: 1) Acceptance criteria 2) User story 3) Definition of done.

QM plays an important role in every stage of the sprint:

- During the Product Grooming meeting, the QM begins gathering NFR by asking the PO and the development team several questions, such as: the system's scalability when more resources are added; the likelihood of the system performing without failure; and how long data should be retained in the system for reference (this might be a government/national regulation); What are the consequences if the user cannot access the system?
- Prior to Sprint planning: QM works with the PO to complete the product backlog.

QM incorporates NFR testing into the test plan.

- Sprint planning: QM presents and explains to the development team the identified NFRs for each user story.
- Daily meeting: QM tracks NFR work progress and assists the team in overcoming difficulties.

QM assists the team in remembering the importance of the quality attributes.

- Review meeting: While the PO focuses on the FR, the QM examines the NFR.
- Retrospective: QM recommends enhancements to increase product quality. For instance, devote more time to quality attributes such as security.

2) *Manage testing activities*: The SQRum method enhances the classic scrum "whole team" approach, in which

every team member is responsible for quality and every team member is a tester [19]. SQRum proposes to give the responsibility for managing testing tasks to a single member who is the Quality Manager. QM's mission is not to pilot and supervise the SQRum team, but to accompany and coach them to ensure that they have all they need to execute QA activities. His role is to:

- Break down testing activities into several tasks.
- Ensure that the appropriate testing tasks are scheduled during release and iteration planning sessions.
- Assign testing responsibilities to team members so that everyone is aware of what to accomplish.
- Ensure that all testers meet their deadlines for work completion.
- Take notice of test-related challenges and attempt to address them.
- Ensure that each tester has the skills and knowledge necessary to develop and perform the tests required for each user story.
- Employ pair testing to address the skills gaps of the tester.
- Help to estimate the overall test effort and the technical resources needed.

3) *Define QA strategy*: In Sprint Zero (also known as the pre-planning phase of a sprint project), the QM collaborates with the scrum team to develop the QA strategy. However, his responsibilities do not end there because he is also responsible for keeping the QA strategy updated. The QA strategy is used by the quality manager to provide a new tester with an overview of the test process.

4) *Establish the test plan*: Before the sprint planning, the QM asks the PO about what stories will be in the next sprint, so he takes time to understand functionally and technically the requirements, and he starts working on his sprint plan. When the sprint planning comes, he already has an idea about quality attributes, possible issues and dependencies, data creation estimation, and test effort. This raises awareness of potential resource, time, and scope of work constraints confronting testers, as well as risks that must be discussed and addressed. This also allows the PO to reevaluate the level of quality he requires, and how much work should fit within the actual, achievable velocity of the sprint.

The Quality Manager may delegate the preparation of the test plan to a member of the team, but he remains responsible for the veracity of the information included within.

5) *Help the team in expressing its DoD*: The Quality Manager assists the team in creating a common understanding of quality to ensure that each user story makes sense within the context of the product's bigger story. The QM helps the team in formulating its DoD by asking the appropriate questions, such as:

Are functional tests passed?

Is acceptance testing finished?

Are quality attributes considered?

6) *Tracking test tasks and status*: At any point in the sprint, the Quality Manager must be able to quickly determine how much testing work remains on each story and which stories are "done." He must also ensure that no story is "done" until it has been tested at all appropriate levels. This helps him to check if the team is on schedule and to anticipate if there is a story that cannot be completed and he must remove it or ask programmers to help with the testing tasks.

Tracking the number of tests produced, executed, and passed at the story level helps indicate the status of a story. The number of tests written shows the progress of tests to drive development. Knowing how many tests aren't passing yet gives you an idea of how much code still needs to be written. The burndown chart is an example of a method used for measuring team progress.

Story or task boards are a helpful visual way to determine the state of an iteration, particularly if color coding is utilized; there are different colored index cards for the various types of tasks, such as green for testing, white for coding, and yellow and red for defects. Progress tracking can be achieved by any method and with both virtual and physical storyboards, as long as it enables the QM to see at a glance how many stories are "done," with all coding, database, and testing completed, and whatever the team's DoD is.

7) *Identify risks and threats to sprint*: Every user story in the product backlog is a potential risk. A story risk is the level that a user story will fail (the impact of the failure multiplied by the probability of failure). The key purpose of the risk prediction is to accurately anticipate the testing work so that all user stories can be tested in accordance with the risk level defined by the entire team. A simple test is sufficient for stories with a low likelihood of failure. Unlike stories with a high failure risk, which require a very careful test plan containing a variety of test techniques. Stories with a high failure risk require a very careful test plan containing a variety of test techniques. The QM assists the team in achieving the ideal balance between sufficient quality and acceptable risk, on the one hand, and time and resource limits, on the other.

The QM can initiate the identification and assessment of risk for both functional and non-functional requirements prior to sprint planning, and the team can finish the risk analysis during sprint planning. If there is not enough time to address the relevant risks at this meeting, the QM can ask the Scrum Master to organize a risk poker session. Once the risks have been identified, the team classifies and evaluates them based on likelihood and impact. This assessment is recorded in the test plan and taken into account during the design, implementation, and execution of tests for this iteration.

8) *Track defect*: Since quality is the concern of the entire team, everyone works collaboratively throughout. The Quality Manager's responsibility is to assist the SQRum team in setting

targets relating to defects and using the right metrics to assess progress toward these goals. Gathering metrics on defects helps reflect the trend, which means the growing attitude in the number of defects in the defect backlog over a period of time. Another QM's responsibility is to communicate the trend in the defect backlog to the SQRum team. If the defect backlog is decreasing, there are no concerns. If it is increasing, the SQRum team must invest time in analyzing the underlying cause. In order to address the root cause, the QM must tell the SQRum team about the nature of the defects. If the defects could not have been detected using unit tests, then perhaps the programmers need additional training in unit test writing. If defects are missed or functional requirements are misinterpreted, then perhaps not enough effort is spent on sprint planning or acceptance tests are insufficiently thorough.

The QM can use a visual technique such as "Defect Trend chart." As shown in Fig. 5, the "Defect Trend chart" is a graphical representation of reported defects over time. The x-axis represents a period of time, while the y-axis represents the number of defects.

9) *Manage defects backlog*: The Quality Manager is in charge of the Defect Backlog, which includes what's in it, when it's available, and how it's organized. Before sprint planning, the quality manager makes an initial selection of defects based on their severity. Product backlog grooming is an excellent opportunity to discuss this selection with the product owner in order to determine which defects to fix first. The selected defects are presented to the team during the sprint planning and are scheduled for the next sprint.

The QM must ensure that the team does not accumulate technical debt, particularly when working with legacy code. The longer a defect remains in the system, the greater its impact. Defects in a code base have negative effects on code quality, system security and flexibility, team effectiveness, and velocity.

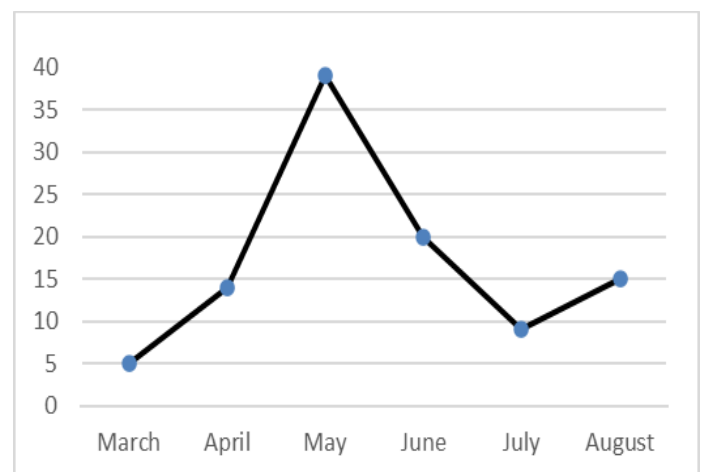


Fig. 5. Defect Trend Chart.



Early detection and correction of defects is more cost-effective. In iSixSigma Magazine, Mukesh Soni [20] cites an IBM report stating that a defect discovered after product release costs four to five times as much as one discovered during product design and up to 100 times as much as one discovered during product maintenance.

The QM must constantly evaluate the amount of technical debt dragging it down and work on reducing and preventing it. The QM needs to persuade the product owner of the benefits of addressing technical debt by demonstrating that technical debt may be costing the business money due to decreasing velocity. The team's velocity is sometimes consumed by bug fixes and trying to make sense of the code.

The QM can request that the PO reduce the scope of his desired features to allow sufficient time for good practices such as continuous small refactoring, which results in improved test coverage, a solid foundation for future development, decreased technical debt, and higher overall team velocity.

If it is insufficient and the PO cannot budget time in each iteration, the QM may suggest to the PO that a "refactoring iteration" be planned as a last resort to upgrade or add necessary tools, reduce technical debt, automate more tests, and perform major refactoring efforts. Planning refactoring iterations at regular intervals improves quality, maintains the system and its infrastructure, and preserves the team's velocity, allowing the team to move faster.

*10) Help the team stay focused on the big picture:* he Quality Manager tries to put each story in the context of the whole system by looking at possible risks, dependencies, and unplanned effects on other parts. The QM assumed the perspectives of the user, product owner, programmer, and tester, as well as everyone engaged in building and using the features. He can consider the effects of FR and NFR on the larger system and bring this to the attention of the team. Everyone on the team may easily focus their attention on the work or story at hand. This is a disadvantage of working on small feature portions at a time. The goal of the QM is to help

the team take a step back and evaluate how their current stories fit into the big picture. QM keeps challenging the team to do a better job of delivering real value.

*11) Keep testing environment updated:* Testers cannot test effectively in the absence of a test-controlled environment. The QM must continuously inspect test environments and collect information regarding the deployed build, database schema, whether or not somebody is altering the schema, and other processes operating on the system. This information enables him to sustain the test environment with the most recent or updated version and eliminate the obsolete test environment, its tools, and techniques. This is also true for databases. Sometimes other teams can modify fields, add columns, or remove obsolete ones. The QM must be aware of all these changes in order to keep his database updated.

#### *D. SQRum Process*

We have identified eleven quality manager responsibilities. This list of eleven responsibilities indicates the tasks for which the QM is accountable; the remaining QA activities can be performed by the rest of the team, since quality is still owned by the entire team (everyone is a tester), but managed by just one person.

SQRum method follows eight phases, which are: Project Initiation, Sprint 0, Product Grooming, Sprint Planning, Sprint Execution, Sprint Demo, Sprint Retrospective, and Release. These phases are described in Table I with the artifacts of each phase.

#### *E. SQRum Metamodel*

This section presents the proposal of the new SQRum method as a metamodel. As shown in Fig. 6, the SQRum Metamodel is based on the transformation of the method's concepts into metaclasses linked by meta-associations, which define the kinds of relationships between these concepts. The green metaclasses represent the new concepts added to the Scrum method related to QA viz, Test Library, Defect Backlog, Definition of Done, QA Strategy, Refactoring Iteration, Test Plan, Product Grooming, and Quality Manager.

TABLE I. SQRUM PROCESS

	Project Initiation	Sprint 0	Grooming	Sprint Planning	Sprint Execution	Sprint Demo	Sprint Restro	Release
<b>Activities</b>	<ul style="list-style-type: none"> <li>Create project Idea</li> <li>Define project start/end dates</li> <li>Team composition</li> <li>Get a Quality Manager</li> <li>Define sprint length</li> </ul>	<ul style="list-style-type: none"> <li>Train the team in the SQRum method</li> <li>Communicate the role of QM</li> <li>Define the QA strategy</li> <li>Setup the test environment</li> <li>Build the test infrastructure</li> </ul>	<ul style="list-style-type: none"> <li>Identify product's quality attributes</li> <li>Identify risk and dependencies</li> <li>Review the future scope</li> </ul>	<ul style="list-style-type: none"> <li>Plan tests</li> <li>Keep testing environment updated</li> <li>Define DoD</li> <li>Identify acceptance criteria</li> <li>Estimate test effort</li> <li>Plan tests automation</li> <li>Identify test data</li> <li>Participate in sizing stories</li> <li>Complete product backlog with NFR</li> </ul>	<ul style="list-style-type: none"> <li>Track tests activities</li> <li>Track defect</li> <li>Report defect</li> <li>Write and execute tests campaigns</li> <li>Perform nonfunctional testing</li> <li>Communicate tests results</li> <li>Report test impediments</li> <li>Create test data</li> <li>Run automated testing scripts</li> <li>Automate new functional tests</li> <li>Review of resolved defects</li> <li>Pair-test with other testers</li> </ul>	<ul style="list-style-type: none"> <li>Check functional and non-functional requirements</li> <li>Report defect</li> </ul>	<ul style="list-style-type: none"> <li>Inspect the process and people</li> <li>Identify improvements</li> </ul>	<ul style="list-style-type: none"> <li>Participate in release to production</li> <li>Train end users</li> </ul>
<b>Artifacts</b>	Project idea	QA strategy	Product backlog Defect backlog	Product backlog Sprint backlog DoD Defect backlog Test plan	Sprint backlog Increment Defect backlog Test library	Increment	QA strategy Test plan	Increment

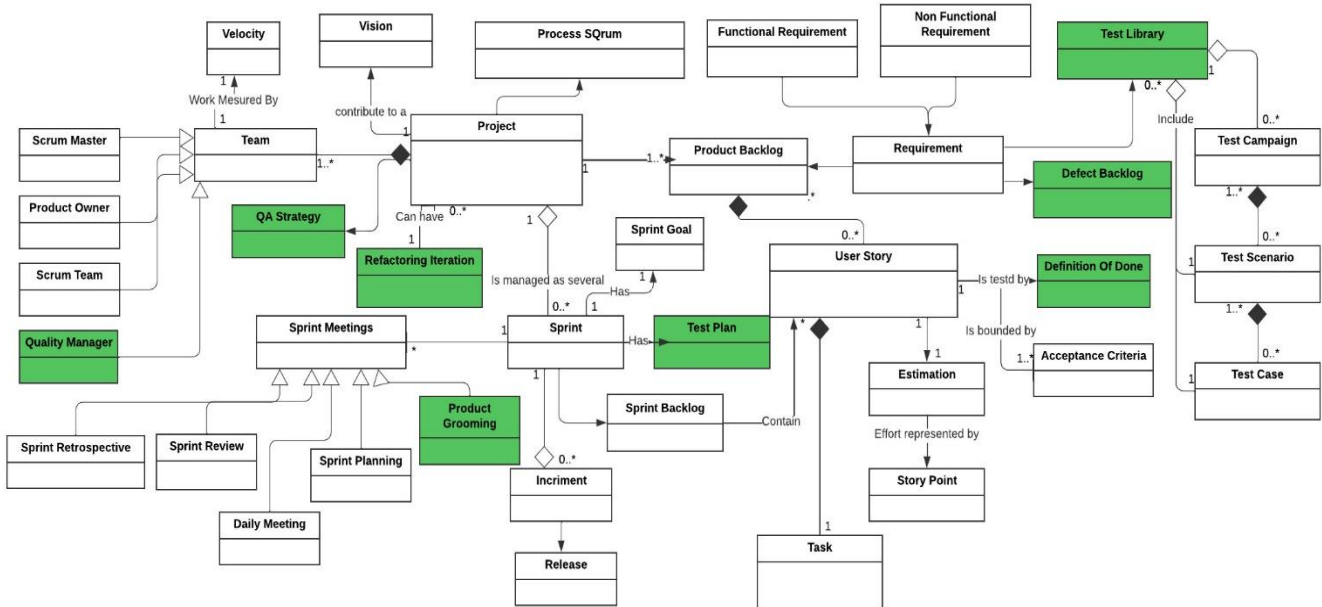


Fig. 6. Proposed Metamodel for SQRum Method.

## VI. CASE STUDY AND EVALUATON

### A. Project Description

To validate our model, we conducted a four-month case study on a software development project in a multinational telecommunications and IT services company, which is a large

mobile network operator that serves European and worldwide companies. When we began the case study, the project had been underway for more than a year, and the team was using the agile scrum methodology. The main goal of this project is to rebuild a legacy system used by the project managers to plan and manage the deployment of internet solutions for enterprise

customers (btob business model). By rebuilding the system, the company wants to establish a new technological system, standardize further development and maintenance procedures, improve the system's quality and facilitate project managers' work. The team has agreed to test our method to determine if it will help them achieve this goal. Before starting the use trial, the team received SQRum training. The case study was distributed into four increments. Three are development sprints, and one is a refactoring sprint. Sprints were 4 weeks long. One quality manager, one product owner, seven developers, and one scrum master made up the team that worked on the project.

The company required a 10-month trial period, and if the deployment is successful, it will adopt SQRum as a standard methodology for two of its projects.

### B. First Results

This section presents results after only four months of using SQRum. It is a classification of the team's feedback about which aspects they think have been improved using SQRum.

1) *Awareness about quality*: All the team members reported that SQRum made them more aware of software quality during the sprint. Developers paid more attention to quality and were more focused.

2) *Better organization and more communication*: Communication and organization were also cited by the team as SQRum improved aspects. Since the QM took over the management of everything related to quality, the Scrum Master was able to focus on monitoring development activities and assisting the team in overcoming obstacles.

The team was also able to communicate more effectively as a result of the new SQRum events, since each event provided an opportunity for sharing and interacting.

3) *Efficiency improvement*: The team members reported improved efficiency. By mapping out the functional requirements and quality attributes, the QM helps the PO define clearer user stories from a technical perspective. Due to the clarity of the user stories, the team was more efficient and had a better knowledge of what to do and how to do it, as well as a better understanding of potential risks, which has improved not only the efficiency of the team but also the quality.

4) *Better use of team velocity*: Analysis and verification of quality attributes on a periodic basis, as well as the implementation of a refactoring sprint, led to a reduction in defect counts and defect-fixing time, resulting in a 21% increase in development productivity. The team spent more time in providing value rather than fixing issues.

5) *Testing properly*: The team found that the tests had been enhanced; they were better documented and structured as a result of the use of the test library. That helped them keep the balance between communication and documentation. Developers also reported that with the help of the test plan, they were able to test more thoroughly and were aware of the

aspects to be taken into account when testing in order to improve quality.

6) *Improving quality*: Initiating the refactoring iteration, checking the quality attribute, and using the newly proposed artifacts decreased the defect density and time required to fix defects. The prevention and control of bugs contributed to a 36% decrease in the defect density of the project, and the time necessary to remedy defects was decreased by 41% by easily locating the spot of change and estimating side-effects.

### C. Aspects to be Improved

Despite the positive results of SQRum, months is not enough time to test all its aspects. Also, for a better assessment of quality, the method needs to be used on large-scale projects with SAFe.

The members suggested detailed guidance on analyzing the quality attributes and enhancing the traceability, they also suggest a burn-down chart to assess the current state of the QC activities in SQRum.

## VII. DISCUSSION AND CONCLUSION

Scrum is the most used method because it is adaptable to all project types. It is an iterative and incremental method that helps teams to deliver a high-quality project. Scrum's primary issues continue to be process and product quality.

This study introduces SQRum which is an adaptation of Scrum that includes and promotes the existence of a quality owner role.

SQRum provides a quality enhancement by adapting the traditional Scrum to emphasize non-functional requirements, establishing a new role and new artifacts to focus on control assurance activities, and ensuring that the quality assurance process has been adhered to.

In an ideal world, the goal would be to have zero defects, but due to the sprint's short lifecycle, this goal is almost impossible to achieve. Quality should not be seen as a constraint, but rather as a tool for maximizing business value. The QM's responsibility is to assist the PO in finding the ideal.

### REFERENCES

- [1] T. Khalane and M. Tanner, "Software quality assurance in Scrum: The need for concrete guidance on SQA strategies in meeting user expectations," in 2013 International Conference on Adaptive Science and Technology, Pretoria, South Africa, Nov. 2013, pp. 1–6. doi: 10.1109/ICASTech.2013.6707499.
- [2] A. Srivastava, S. Bhardwaj, and S. Saraswat, "SCRUM model for agile methodology," in 2017 International Conference on Computing, Communication and Automation (ICCCA), Greater Noida, May 2017, pp. 864–869. doi: 10.1109/CCAA.2017.8229928.
- [3] L. Crispin and J. Gregory, *Agile Testing: A Practical Guide for Testers and Agile Teams*. Pearson Education, 2009.
- [4] "MetaObject Facility | Object Management Group." <https://www.omg.org/mof/> (accessed Sep. 01, 2022).
- [5] G. K. Hanssen, B. Haugset, T. Stålhane, T. Myklebust, and I. Kulbrandstad, "Quality Assurance in Scrum Applied to Safety Critical Software," in *Agile Processes, in Software Engineering, and Extreme Programming*, vol. 251, H. Sharp and T. Hall, Eds. Cham: Springer International Publishing, 2016, pp. 92–103. doi: 10.1007/978-3-319-33515-5\_8.

- [6] S. Jeon, M. Han, E. Lee, and K. Lee, "Quality Attribute Driven Agile Development," in 2011 Ninth International Conference on Software Engineering Research, Management and Applications, Baltimore, MD, USA, Aug. 2011, pp. 203–210. doi: 10.1109/SERA.2011.24.
- [7] O. P. Timperi, "An Overview of Quality Assurance Practices in Agile Methodologies," p. 10, 2004.
- [8] M. Aamir and M. N. A. Khan, "Incorporating quality control activities in scrum in relation to the concept of test backlog," *Sādhanā*, vol. 42, no. 7, pp. 1051–1061, Jul. 2017, doi: 10.1007/s12046-017-0688-7.
- [9] N. Bajnaid, R. Benlamri, and B. Cogan, "An SQA e-Learning System for Agile Software Development," in *Networked Digital Technologies*, vol. 294, Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 69–83. doi: 10.1007/978-3-642-30567-2\_7.
- [10] K. Schwaber and J. Sutherland, "The Scrum Guide."
- [11] K. Schwaber, "SCRUM Development Process," p. 18.
- [12] M. Soukaina, E. Badr, M. Abdelaziz, and S. Nawal, "Towards a New Metamodel Approach of Scrum, XP and Ignite Methods," *IJACSA*, vol. 12, no. 12, 2021, doi: 10.14569/IJACSA.2021.0121225.
- [13] R. Pichler, *Agile Product Management with Scrum: Creating Products that Customers Love*. Addison-Wesley Professional, 2010.
- [14] F. Ramos, A. A. M. Costa, M. Perkusich, H. Almeida, and A. Perkusich, "A Non-Functional Requirements Recommendation System for Scrum-based Projects," Jul. 2018, pp. 149–187. doi: 10.18293/SEKE2018-107.
- [15] ISO/IEC 9126-1:2001, "Software engineering – Product quality – Part 1: Quality model."
- [16] A. Silva et al., "A systematic review on the use of Definition of Done on agile software development projects," in *Proceedings of the 21st International Conference on Evaluation and Assessment in Software Engineering*, Karlskrona Sweden, Jun. 2017, pp. 364–373. doi: 10.1145/3084226.3084262.
- [17] R. Moser, P. Abrahamsson, W. Pedrycz, A. Sillitti, and G. Succi, "A Case Study on the Impact of Refactoring on Quality and Productivity in an Agile Team," in *Balancing Agility and Formalism in Software Engineering*, vol. 5082, B. Meyer, J. R. Nawrocki, and B. Walter, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 252–266. doi: 10.1007/978-3-540-85279-7\_20.
- [18] F. Ribeiro, A. L. Ferreira, A. Tereso, and D. Perrotta, "Development of a Grooming Process for an Agile Software Team in the Automotive Domain," in *Trends and Advances in Information Systems and Technologies*, vol. 745, Cham: Springer International Publishing, 2018, pp. 887–896. doi: 10.1007/978-3-319-77703-0\_86.
- [19] S. Najihi, S. Elhadi, R. A. Abdelouahid, and A. Marzak, "Software Testing from an Agile and Traditional view," *Procedia Computer Science*, vol. 203, pp. 775–782, 2022, doi: 10.1016/j.procs.2022.07.116.
- [20] S. Mukesh, "Defect Prevention\_ Reducing Costs and Enhancing Quality," p. 6.

# Use of Interactive Multimedia e-Learning in TVET Education

Siti Fadzilah Mat Noor<sup>1</sup>, Hazura Mohamed<sup>2</sup>, Dayana Daiman<sup>4</sup>

Faculty of Information Science and Technology  
Universiti Kebangsaan Malaysia  
43600 Bangi, Selangor, Malaysia

Nur Atiqah Zaini<sup>3</sup>

School of Computing and Creative Media  
University of Technology Sarawak  
96000 Sibu, Sarawak, Malaysia

**Abstract**—Malaysia is focused on the development and use of technologies among consumers. Thus, technological innovations are used in the adaptation of online learning to educate students, as well as to enhance the teaching and learning process in Technical and Vocational Education and Training (TVET) institutions. There is a need to expose students to the online learning revolution, which conceptualises using computerised systems to facilitate the learning process. However, the COVID-19 outbreak has disrupted the academic year across the country. Due to the unusual circumstances related to the pandemic, the Malaysian government has urged all academic institutions to conduct online teaching and learning. Thus, an e-Learning system, known as SpmiILP, has been designed and developed accordingly for an interactive multimedia course to encourage online interaction among students and lecturers, as well as to enhance human learning and cognitive development. In fact, these essential elements such as learning style of the students and user experience are focused on to engage them in learning effectively as well. An e-Learning System for Interactive Multimedia Course was used to develop the e-Learning system (SpmiILP). The usability test showed that the developed e-Learning system has a positive influence that provided potential contributions to (TVET) students in their learning processes.

**Keywords**—e-Learning; interactive multimedia; learning style; user experience

## I. INTRODUCTION

The evolution of information technology has spread worldwide in conjunction with the rapid growth of various technologies. Multimedia technology has demonstrated the great potential of evolution in learning, accessing, and manipulating information. Multimedia contributions have an enormous opportunity for educators to expand numerous learning techniques, especially e-learning. e-Learning shows the potential of digital transformation in the education system that would allow the development of new teaching and learning ecosystems. e-Learning growth was facilitated using digital tools involving interactivities that encourage online interaction between students and lecturers. e-Learning concepts encompass many multimedia technologies and the Internet that enable access to virtual learning environments. The use of information and digital technology empowers the diverse learning process by combining traditional classroom and online learning environments. Therefore, e-Learning has been massively implemented in higher education institutions based on pivotal success factors categorised as system quality,

service quality, information quality, usefulness, and engagement.

The advancement of teaching and learning technologies has helped improve students' critical thinking skills. e-Learning's enormous growth, alongside technological advancement, has promoted high-quality learning techniques that fit students' preferences. The term "digitally literate" indicates the ability to perform and handle digital technologies humans use daily [1],[2]. Indeed, the quality of learning has been enhanced by advancing educational tools. Nowadays, younger generations are immersed and surrounded by technologies, with easy access to information.

The e-Learning approach supports collaborative communication that allows users to control and customise their learning environment. Satisfaction in e-Learning has enabled users to experience learning that fits their preferences and styles [3]. Multimedia elements also play a significant role in education by allowing users to experience interactive learning. Multimedia elements consisting of text, audio, animation, image, and video can trigger users to be more entertained and keener to continue learning [4]. The intensive penetration of multimedia has fuelled the demand for visualising educational materials to be more engaging and effective.

Hence, this next section of the paper will further provide detail on the related works; methodology will emphasize the five phases of analysis, design, development, implementation, and evaluation, then will summarize the results, discussion and conclusion.

## II. RELATED WORKS

The massive spread of coronavirus disease or COVID-19 has become a global pandemic that forces social distancing policy. Progressive steps are taken to limit the spread of this policy in the community and influence various sectors, including education. Therefore, there is a sudden transition of learning where the institutions are required to implement online learning. Human motion restrictions force changes in our education in new aspects such as learning styles, learning platforms, accessibility, and the deliverance of information [5]. Thus, the pandemic of COVID-19 leads educators to provide learning materials through online learning. The transition from face-to-face learning to remote learning as an alternative shows the potential of e-Learning that uses an online platform. To that end, face-to-face learning has become vulnerable in most

institutions due to the COVID-19 pandemic, which results in another alternative of providing online learning for the students.

In a world with ever-evolving computing technology, e-learning has been pushing the advanced technology boundaries. e-Learning is recognised as an important part of learning in higher education institutions. The expansion of e-learning has initiated several changes in education delivery, as shown in Fig. 1 [6].

This e-learning model shows the different e-learning techniques in education that learners can employ. First, adjunct e-learning can be used in a traditional classroom that provides relative independence to learners. Second, the blended e-Learning technique can explain the delivery of course materials using traditional learning and e-Learning methods. The third technique is conducted online, devoid of traditional learning or classroom participation, because it features individualised and collaborative learning. The development of the proposed e-learning system in this research paper was focused on blended e-Learning.

Therefore, Industrial Training Institutes, or Institut Latihan Perindustrian (ILP), provide formal training for school leavers and industrial workers that would enable them to acquire skills for specific work fields. By upgrading their skills, they would be more focused on their work and contribute effectively to this country's development. Interactive Multimedia has been introduced to ILP students as one of the main subjects in the Information Technology (IT) course. ILP students have been using the conventional way of learning instead of focusing on the rapid growth of digital technology for teaching and learning [7]. The lack of motivation and engagement has caused them to lose interest in learning.

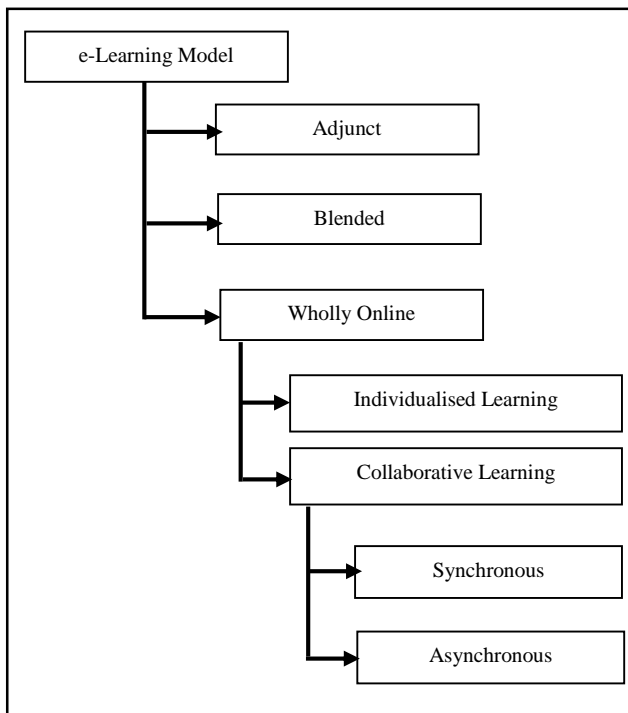


Fig. 1. e-Learning Model.

Nonetheless, the ongoing technological changes have enhanced the education system in this country with the use of digital technologies. The e-Learning system for the Interactive Multimedia subject can drive ILP students to use technology and learn effectively. e-Learning has become a learning platform with various benefits, such as being easy to use, user-friendly, interactive, and efficient for students [8].

Online learning has been introduced in ILP to expose students to different teaching and learning methods that would suit their learning styles. Some students prefer to study in groups, while some prefer to study alone. Additionally, some of the students' performances showed that they understand better through verbally explained lessons, while some might have difficulties in this environment. They might prefer learning using visual and graphical forms of the lesson to understand and improve their motivation and engagement [9],[10]. These different learning styles show that students have different capabilities in adapting to knowledge. Thus, the e-Learning system for the Interactive Multimedia subject has been developed with several functionalities, such as interactive notes, videos, and quizzes. These functionalities enable the students to complete the tasks involving problem solving and assignments. The students can improve their cognitive abilities by adopting an online platform during the pandemic COVID-19. The user interface is obtained based on the user requirements specified on the usage of multimedia elements such as animation, text, videos, and audio. Using multimedia elements helps them stay focused and have fun while learning.

The students involved in this study could complete their assessments and were evaluated right after they completed the e-Learning. It should be aligned with the users' performance and learning styles to make the e-Learning system more effective. The constantly growing research on interactive multimedia e-Learning systems has helped foster new approaches to learning. Therefore, this study has aimed to design and develop an e-Learning system for the Interactive Multimedia subject (SpmiILP) for ILP students that could improve their learning outcomes.

In the meantime, students can utilise other currently available systems, such as Moodle, as a learning management system. Moodle offers various features, such as reading materials, papers and projects, forums relating to the course, conduction of quizzes, the distribution, collection, and evaluation of assignments, keeping track of class attendance, and recording grades. Some of these features have been used to develop the e-Learning system (SpmiILP) based on the user requirements of ILP students enrolled in the Interactive Multimedia subject.

The SpmiILP was adapted to fit the needs of individual learners. For example, it offers students the flexibility of time and place to access the huge amount of information in the system, eliminates the barriers that could potentially hinder student participation in the classroom, and it enables the students to study at their own pace and speed [6], [7],[11]. As a result, they can be satisfied with their performance and decrease their stress level. Since every student's learning style must be taken into consideration, this paper presents the design



and development of an e-Learning system that is focused on the user requirements of ILP students.

### III. METHOD

This study was conducted in five phases: analysis, design, development, implementation, and evaluation. Each phase is explained in detail in the following sections.

#### A. Analysis Phase

This phase involved analysing user requirements obtained from the users, as shown in Table I. Previous studies have highlighted the learning styles of undergraduate students, multimedia elements, user interaction, user experience, usability, and types of technology used [7], [8], [9], [11], [12], [13]. Identifying user requirements of a system has helped strengthen their understanding of the learning process.

The user requirements listed in Table I have been acquired to develop an e-Learning system for the Interactive Multimedia subject. Instead of focusing on interface designs, this e-learning system was developed to meet students' requirements with functional features that could engage and motivate them in their learning process. ILP students mostly prefer learning using different forms of visual and graphic learning tools, including animation, text, video, and audio. The tasks given to the students were in the form of online learning by providing quizzes and assessments. The implementation of online learning has enhanced the use of technologies among these students that they can use to obtain unlimited access to information.

#### B. Design Phase

This phase involved utilising the outputs obtained from the analysis of user requirements. The selection of activities must be suitable for ILP students to improve their learning performance. This system has adapted multimedia elements that showed greater potential and have gained popularity among the students. Fig. 2 shows a model of the e-Learning system for Interactive Multimedia subject (SpmiILP). This model consists of user specifications and software content.

TABLE I. USER REQUIREMENTS OF ILP STUDENTS

User Requirements	
1.	Device Used: Computer/Laptop
2.	User Interaction
3.	Learning Styles: Visual & Graphics
4.	Multimedia Elements: Animation, Text, Video, Audio
5.	Usability: Effectiveness
6.	User Experience: Engagement, motivation
7.	Activities: Quizzes & Assessments

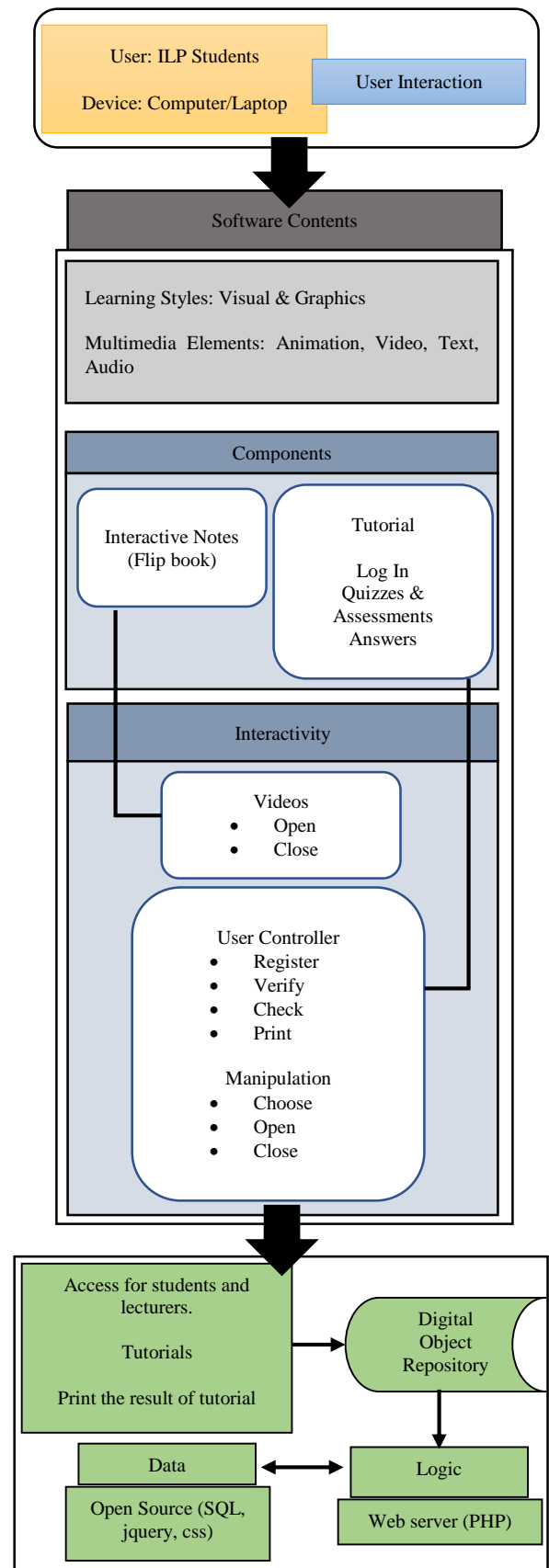


Fig. 2. Model of the e-Learning System for Interactive Multimedia Course (SpmiILP).

The target users of this system were the ILP students, who interactively used current technology, such as computers or laptops. The software contents of this system consisted of different learning styles: visual and graphical learning tools; multimedia elements, such as animation, video, text, and audio; and other components, such as interactive notes (Flipbook) and tutorials. This flipbook consisted of several learning materials, such as notes, videos, and animation that can enhance the learning experience using technology. The interaction between the users and the system occurred through buttons; for example, videos were provided with open and close buttons. Meanwhile, accessing the tutorials with functionalities that consisted of login, quizzes, assessments, and answers has enabled the users to choose, open, close, register, verify, check, and print the results using these buttons. PHP MySQL was the database system for SpmiILP, and HTML was used for the interface design template to develop this e-learning system.

Fig. 3 depicts the hierarchy of the e-learning system's interface design, which includes login, user ID, password, notes, upload notes, list of notes, students, registration, list of students, and quizzes.

C. Development Phase

Fig. 4 shows the interface design for the SpmiILP that ILP students use to access and study the Interactive Multimedia subject. This is the main page for student registration, and this interface consists of the main page, student registration, upload notes, interactive notes, quizzes, and results. Before logging in, the students need to obtain their user ID and password by filling in their details, such as matric number (NDP), name, gender, session, and phone number. These details are compulsory for registration, where the ID and password are set up the same as the NDP for the students. The main page of this e-Learning system consists of the main menu, notes, students' names, lists of quiz categories, quiz sections, and results.

Fig. 5 shows the feature for uploading notes in the e-learning system. This page can be accessed by both the lecturers and the students. The lecturers can upload their visual and graphical notes for teaching and learning. Instead of learning by reading notes consisting only of texts, this method could help students engage more in the learning process.

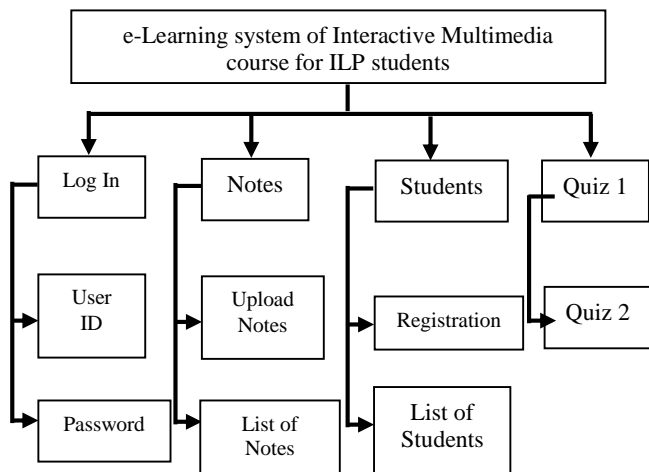


Fig. 3. Hierarchy of the e-Learning System Interface Design.

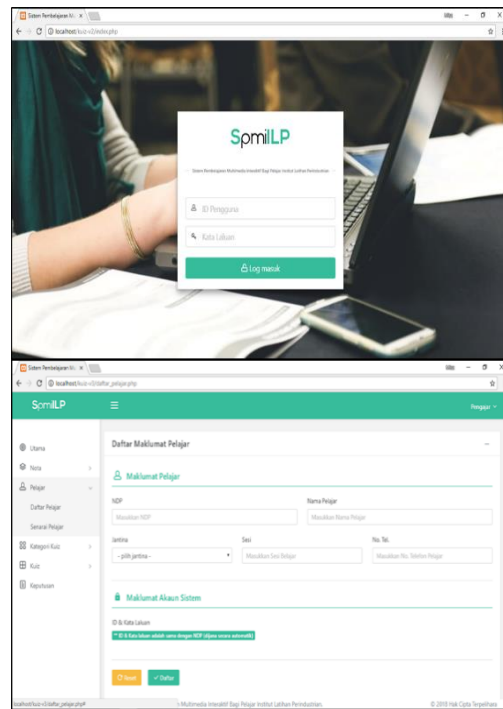


Fig. 4. Students' Registration.

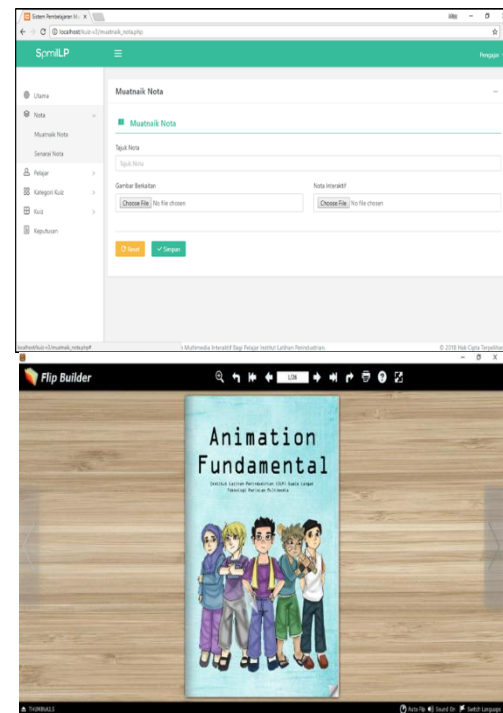


Fig. 5. Features in the e-Learning System (Uploading Notes and Interactive Notes).

Fig. 6 shows the interface for the quizzes section. In this section, the lecturers can update the students' quizzes list. The quizzes section aimed to evaluate students' understanding of the subject they were learning. It can also provide lecturers with insights into students' progress. After the lecturers have finalised the quizzes, the students can answer them within the estimated time.

Fig. 7 shows the interface for quiz results. The students can check their marks after completing their quiz sessions. They would be evaluated based on their performance in answering the quizzes. The lecturers can access this page, and the students can view their grades. This feature showed the user interaction among the students through this platform. Therefore, the lecturers could use this platform to impart to their students a great deal of knowledge.

D. Implementation and Evaluation Phase

The usability of the SpmiILP was evaluated among ILP students using questionnaires. There were 24 respondents from the Interactive Multimedia class in this usability evaluation. Data are obtained from a questionnaire after respondents use the application. The questionnaires consisted of Section A and Section B, where Section A was focused on the respondents' background, as shown in Table II. Meanwhile, Section B focused on three dimensions of the usability evaluation: efficiency, functionalities, and effectiveness. Respondents' responses were based on agreement on all items based on a 5-point Likert scale; namely, 1 = strongly disagree, 2 = disagree, 3 = slightly agree, 4 = agree, and 5 = strongly agree. The usability of this system was analysed using answers scales of 4 and 5 points to get the percentage using Microsoft Excel.

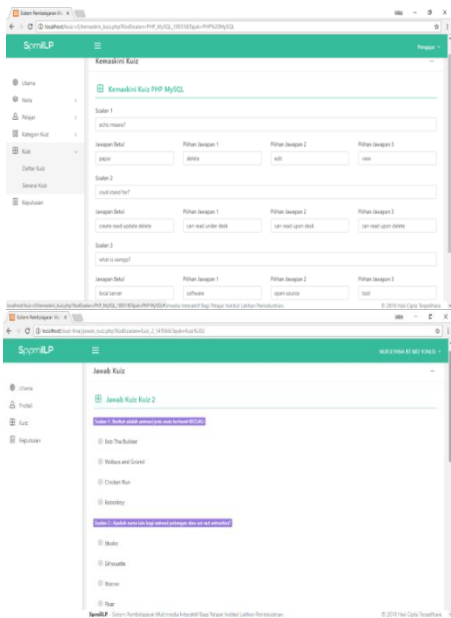


Fig. 6. Features in the e-Learning System (Quizzes Section).

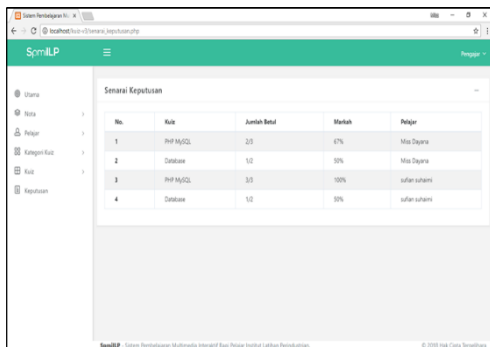


Fig. 7. Features in the e-Learning System (Results).

Table II shows the respondents' background in gender, age, race, education, and application usage in teaching and learning. Based on the collected data, 91% of the respondents had not used the system or other applications during the learning process. This result was troublesome because the development of technologies in Malaysia is growing rapidly, whereby the evolvement of technological innovation can enhance the teaching and learning process. Yet, these respondents have failed to take this opportunity. Therefore, there is a need to develop online learning systems or applications that could enhance the effectiveness of e-Learning and increase students' motivations. By providing the online learning platform to TVET students, they can use it to improve their communication with their educators instead of only focusing on conventional learning.

Fig. 8 shows the percentage of all items in the efficiency dimension. More than 70% of respondents answered using scales 4 and scale 5. The highest percentage was 91% of respondents who agreed with using text, graphics, audio, and videos in the system, which would engage them to learn more. This agrees with another study that reported that multimedia elements attracted the respondents to explore more of the e-Learning system (Nauman et al., 2020). Meanwhile, the lowest percentage was 77% of respondents who agreed with the notes provided. Based on the questionnaire, several problems were found, such as the bland design of the interactive notes and the size of letters that have caused difficulties in reading the notes.

TABLE II. SECTION OF QUESTIONNAIRES (BACKGROUND OF RESPONDENTS)

Details of Respondents	Backgrounds	Total Percentage of Respondents (%)
Gender	Man	77.3
	Woman	22.7
Age	18 – 20 years	100
Race	Malay	95.5
	Indian	4.5
Education	Certificate	100
Have/Have not used system/application in the process of teaching and learning	Yes	9
	No	91

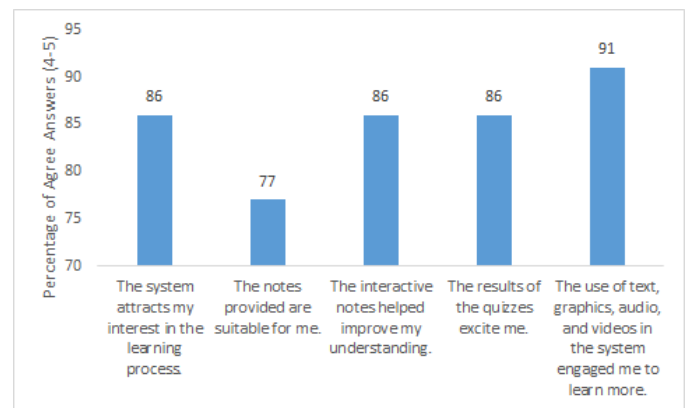


Fig. 8. Efficiency Data Distribution for the e-Learning System.

#### IV. DISCUSSION

Fig. 9 shows that the percentage of all items in the functionality dimension is more than 60%. The highest percentage was 91% (the system was useful for the respondents), and the lowest was 68% (the system was functional, as expected). Therefore, system functionality is the primary determinant supporting students' learning satisfaction. The ability of the system to meet users' requirements has contributed to the overall satisfaction with the system's usefulness [3].

Fig. 10 shows that the percentage of all items in the effectiveness dimension is more than 50%. The highest percentage was 91% (this system is easy to use), and the lowest percentage was 55% (I do not notice any inconsistencies when using this system). Feedback from the respondents showed that inconsistencies occur because of the unstable LAN internet network provided at ILP Kuala Langat.

Overall, the newly developed e-Learning system has helped improve students' performances in the learning process using technology. This system can be used for teaching, whereby the lecturers can upload interactive notes and evaluate the student's performance online. Using multimedia elements and interactive notes have engaged the ILP students in learning.

Developing technologies as essential skills development and learning tools have proved that e-Learning is an effective learning method. The demand for adopting technology-based education, such as e-Learning, has shown an increase in a similar usage, enabling communication with the educators [14], [15]. Instead of only focusing on teaching, combining conventional and technology-based education could improve students' interests. This is because some students could face difficulties articulating their thoughts and would keep the problems to themselves. Through online learning, the educators could help these students figure out their problems [16].

The e-Learning system is an alternative learning system that the students can access online. Due to the advancement of technology, students would easily access the learning materials. The e-Learning system was developed and presented on the platform of mobile devices. Based on the preliminary study, 91% of respondents had not used the online learning system for their teaching and learning process. This has led to several problems, such as the one-way communication between the lecturer and students. The students could also lose their focus during a learning session, and the learning styles can affect their performances [7], [17]. There is a small number of studies conducted on TVET students. Students from the Z generation, or digital native generation, must change their interaction with high technology devices. The e-Learning system emphasises learning activities, social communication, activity control, notifications, file storage, and data security to help the students experience engagement and learnability from the teaching and learning process.

The intensive penetration of digital technology into everyday life has fuelled a new demand in teaching and learning. The educational materials were transformed from conventional learning's functional capabilities into online learning. Adopting online learning, in line with digital technology, can affect the student's learning styles. Their capabilities to learn through online learning have improved their motivation, whereby they could stop being worried or anxious to talk freely with the lecturers. Online learning has helped them learn effectively, whereby the combination of students' learning styles, such as active, visual, and audio, has caused them to be more engaged in the learning process [18].

Consequently, online learning has provided flexibility and mobility because the learning process does not remain at a fixed location and is accessible everywhere. Therefore, the growing outbreak of COVID-19 has forced all academic institutions to adopt online learning. All classes and examinations were cancelled; thus, the sudden shift away from learning has affected the students and lecturers [19]. In response to the significant demand for online learning, institutions of higher learning across Malaysia have provided online platforms to conduct classes and examinations. This shift showed that digital technology would help the users improve their learning motivation. In terms of teaching and learning, the Malaysian education system has changed with the remarkable rise of e-Learning undertaken remotely via digital platforms.

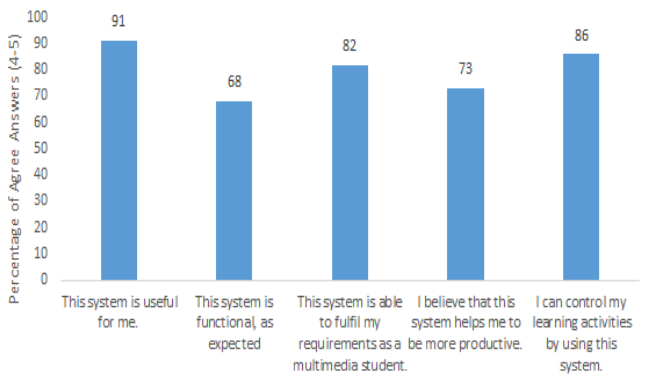


Fig. 9. Functionality Data Distribution for the e-Learning System.

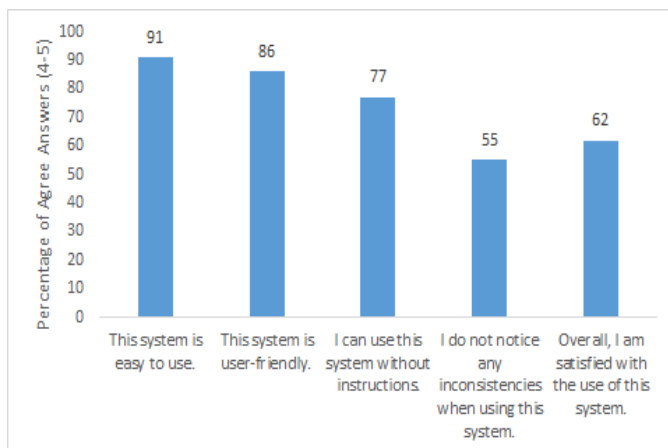


Fig. 10. Effectiveness Data Distribution for the e-Learning System.

In conclusion, the e-learning system for the Interactive Multimedia subject for ILP students has enhanced their performances, motivation, and learning engagement. Different learning styles have caused them to explore various learning methods and improve communication between the lecturers and students. The developed e-learning system has shown the effectiveness and efficiency of this platform and provided functionalities that have enabled the students to access the system easily. The adaptation of the e-learning system has allowed the students to access information at anytime and anywhere. Therefore, e-learning has enhanced the abilities of students to foster the development of digital technology skills. However, the challenges faced in implementing teaching and learning still require the support of educational institutions. Embedding technologies in the classroom help in enhancing teaching and facilitating learning. Broadly, the potential to widen the access and the advancement quality of education need the adoption of technologies that will remain as challenges not only for students but educators as well.

## V. CONCLUSION

The use of technology for teaching and learning has helped improve the quality of the student's performances. The rapid growth of digital technology has allowed educators to utilise various teaching techniques based on the different learning styles among students. Hence, developing the e-learning system for the Interactive Multimedia subject has improved ILP students' engagement in the learning process. The students have adapted well to the functionality and interface design of the system. Several learning techniques have encouraged the students to explore the usage of technology and improve their skills. The developed e-learning system can display teaching and learning materials through computerised digital technology. The system's main features included uploading notes, interactive notes, and quizzes for students and lecturers.

This study has found that the e-learning system has benefited the users in learning interactive multimedia subjects. The rapid growth of online learning methods was due to the advancement of learning technologies and students' needs. With the advancement of the teaching and learning method, the students showed their skills and great potential in mastering other learning methods instead of focusing on conventional learning.

## ACKNOWLEDGMENT

This work is supported by Universiti Kebangsaan Malaysia under the university research grant with grant code: UKM-TR2022-01 and the Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia

## REFERENCES

- [1] Salloum, S. A., Qasim Mohammad Alhamad, A., Al-Emran, M., Abdel Monem, A., & Shaalan, K., "Exploring students' acceptance of e-learning through the development of a comprehensive technology acceptance model", *IEEE Access*, vol. 7, pp. 128445–128462, 2019.
- [2] Lee, J., Song, H. D., & Hong, A. J., "Exploring factors, and indicators for measuring students' sustainable engagement in e-learning", *Sustainability (Switzerland)*, vol. 11(4), 2019.
- [3] Al-Fraihat, D., Joy, M., Masa'deh, R., & Sinclair, J., "Evaluating E-learning systems success: An empirical study", *Computers in Human Behavior*, vol. 102, pp. 67-68, 2020.
- [4] Nauman, A., Qadri, Y. A., Amjad, M., Zikria, Y. Bin, Afzal, M. K., & Kim, S. W., "Multimedia internet of things: A comprehensive survey", *IEEE Access*, vol. 8, pp. 8202-8250, 2020.
- [5] Gherheş, V., Stoian, C. E., Fărcaşiu, M. A., & Stanici, M., "E-learning vs. Face-to-face learning: Analysing students' preferences and behaviors", *Sustainability (Switzerland)*, vol. 13(8), 2021.
- [6] Perrin, D. G., Perrin, E., Brent, M., & Betz, M., "The role of e-learning, advantages and disadvantages of its adoption in higher education", *International Journal of Instructional Technology and Distance Learning*, vol. 12(1), pp. 34-36, 2015.
- [7] Azmi, S., Mat Noor, S. F., & Mohamed, H., "A proposed model of e-learning for technical and Vocational Education Training (TVET) students", *Journal of Theoretical and Applied Information Technology*, vol. 95(12), pp. 2803–2813, 2017.
- [8] Siddiqui, S. T., Alam, S., Khan, Z. A., & Gupta, A., "Cloud-Based E-Learning: Using Cloud Computing Platform for an Effective E-Learning", *Advances in Intelligent Systems and Computing*, vol. 851, 2019.
- [9] Ofosu-Asare, Y. A. W., Essel, H. B., & Bonsu, F. M., "E-Learning Graphical User Interface Development Using the Addie Instruction Design Model and Developmental Research: The Need to Establish Validity and Reliability", *Journal of Global Research in Education and Social Science*, pp. 78-83, 2020.
- [10] Murshed, M., Dewan, M. A. A., Lin, F., & Wen, D., "Engagement detection in e-learning environments using convolutional neural networks", *Proceedings - IEEE 17th International Conference on Dependable, Autonomic and Secure Computing, IEEE 17th International Conference on Pervasive Intelligence and Computing, IEEE 5th International Conference on Cloud and Big Data Computing, 4th Cyber Science*, pp. 80-86, 2020.
- [11] Khairani, N. A., Rajagukguk, J., & Derlina., "Development of Moodle E-Learning Media in Industrial Revolution 4.0 Era", *384(Aistee)*, pp. 752-758, 2020.
- [12] Hadullo, K., Oboko, R., & Omwenga, E., "Factors affecting asynchronous e-learning quality in developing countries university settings", *International Journal of Education and Development Using ICT*, vol. 14(1), pp. 152–163, 2018.
- [13] Hoerunnisa, A., Suryani, N., & Efendi, A., "The Effectiveness of the Use of E-Learning in Multimedia Classes To Improve Vocational Students' Learning Achievement and Motivation", *Kwangsan: Jurnal Teknologi Pendidikan*, vol. 7(2), pp. 123-137, 2019.
- [14] Mustafa, N., Mohd Nordin, N., Embi, M. A., & Norman, M. H., "Testing the Usability of a Mobile Learning Module", *International Journal of Engineering & Technology*, vol. 7, pp. 113-117, 2018.
- [15] Mazin, K. A., Norman, H., Nordin, N., & Ibrahim, R., "Student Self-Recording Videos for TVET Competency in MOOCs Student Self-Recording Videos for TVET Competency in MOOCs", *Journal of Physics: Conference Series*, vol. 1529, 2020.
- [16] Salleh, D., Khairudin, N., Muhammad, F., & Khairudin, R., "Enhancing Social And Lifelong Learning Skills Through The Use Of Mobile Technology As A Motivational Factor", *Journal of Physics A: Mathematical and Theoretical*, vol. 34(1), pp. 17–34, 2020.
- [17] Subramaniam, T. S., Yunus, M. M., Ayub, A. F. M., Rosli, M. S., Maaruf, S. Z., Nawi, A., & Palpanadan, S. T., "Important elements for a framework in designing a mobile learning for english language listening and speaking skills", *Journal of Critical Reviews*, vol. 7(6), pp. 312-315, 2020.
- [18] Mohd Ghazali, A. S., Mat Noor, S. F., & Mohamed, H., "E-Hospitality and Tourism Course Based on Students' Learning Styles", *Asia-Pacific Journal of Information Technology and Multimedia*, vol. 10(01), pp. 100 – 117, 2021.
- [19] Md Ali, S., Khan, S. M. N., Yaacob, N. I., & Baharudin, H., "The Effectiveness of Systems and Applications Usage for Online Final Examination During Movement Control Orderperiod", *Asia-Pacific Journal of Information Technology and Multimedia*, vol. 10(01), pp. 118–123, 2021.



# CBT4Depression: A Cognitive Behaviour Therapy (CBT) Therapeutic Game to Reduce Depression Level among Adolescents

Norhana Yusof<sup>1</sup>

School of Multimedia Technology  
and Communication  
Universiti Utara Malaysia, 06010  
UUM Sintok, Kedah, Malaysia

Nazrul Azha Mohamed Shaari<sup>2</sup>

Faculty of Computer and  
Mathematical Sciences  
Universiti Teknologi MARA, 40450  
Shah Alam Selangor, Malaysia

Eizwan Hamdie Yusoff<sup>3</sup>

Faculty of Medicine  
Universiti Teknologi MARA  
68100 Batu Caves, Selangor,  
Malaysia

**Abstract**—Dropping out of depression treatment commonly occurs in the current psychotherapy treatment. Adolescents often find it difficult to express their thoughts and feelings clearly due to their developmental constraints. They also have trouble realising their behaviours as unhealthy or problematic. The use of therapeutic games in depression treatment among adolescents can enhance the engagement level. Indirectly, the issue of dropping out can be reduced among the adolescents. Therefore, this study aimed to improve engagement levels and reduce depression level among adolescents with depression by designing a therapeutic game. A prototype named CBT4Depression was developed in this study. A quasi experimental study was conducted to evaluate the developed therapeutic game and 115 adolescents were recruited to measure their depression level using CBT4Depression. Based on the findings from the evaluation process, it can be concluded that the CBT4Depression considered success to engage and reduce the depression level among adolescents.

**Keywords**—Therapeutic; game; depression; adolescents; cognitive behavior therapy

## I. INTRODUCTION

In recent years, a markedly increased number of adolescents diagnosed with depression each year throughout the world. According to the World Health Organization (WHO), depression is one of the leading causes of illness and disability among adolescents aged 10-19 years [1]. This is, perhaps, not surprising that the statistics of suicide and homicide also increased in line with the numbers of depression, and have become the third-leading cause of death in 15-19-year-olds as reported by WHO. According to Hink et al., [2], this age group is 4 times more likely to die by committing suicide compared to other older adolescents.

As for the treatment, disturbed adolescents need to undergo a number of sessions in psychotherapy but the limitations in the current talk-based treatment may lead to negative experiences among the adolescents [3]. They also face difficulty expressing their feeling verbally and may also deny that they are suffering from depression [2]. Adolescents with depression also may have limited knowledge regarding depression, such as the differences between normal sadness

and depression, early symptoms and impacts that could affect their daily life. In addition, adolescents also have several barriers to seeking professional help or appropriate services such as financial problems, lack of knowledge in the help-seeking process, and their attitudes towards the psychological treatment [3], [4].

Afterwards, this problem could contribute to the higher possibility of dropping out of the treatment [5]. Being aware of the flaws in conventional psychotherapy, researchers and therapists started utilising the popularity of game technology as an assistive tool, also called therapeutic games. The rapid innovations and advances in information and communication technologies have brought a great positive impact on the gaming industry, especially among the young generation. They become very attached to digital games [6]. Psychotherapists have realised the advantages of using games as assistive tools in psychotherapy among young patients. The use of serious games is already being applied to various types of mental illnesses, such as anxiety, depression, phobia, panic disorder and eating disorder [7].

With the arrival of recent game technologies in this area, it can be seen a rapid proliferation of therapeutic games in psychotherapy practices. The advent of game technology has facilitated psychotherapists to have a better understanding of their patients, particularly adolescences. Furthermore, the level of relationship, or rather, engagement, between therapist and adolescent could be enhanced. One of the most intriguing aspects of therapeutic games is that they could promote patients' engagement and motivation during the treatment, and indirectly enhance the success rate of the therapeutic process.

In the earlier version of digital games, it is only used as entertainment tools. To date, digital or serious games are not only utilised in training but also in education, medical and military simulation. The popularity of serious games has grown extensively and is broadly accepted by various age groups, ranging from children to adults. Given the wide popularity and benefits that can be obtained from using serious games, this has increased the interest of researchers and health professionals in using serious games in treatments as assistive tools [8].

---

This study was funded by the Universiti Utara Malaysia through Journal Publication Fee Funding Scheme (SPYPJ).



Due to this growing demand, new alternative technologies for treatment and therapeutic support of various mental illnesses are being developed and implemented [9], [10]. Despite the broad range of effective treatments available for depression, there is still a need for more research to support their use in clinical treatments [11].

The use of games in healthcare settings has increased these past few years. The present studies also display the effectiveness of therapeutic games as an assistive aid in mental health interventions [12]. Moreover, recent research has shown that using games in psychotherapy can help establish the therapeutic relationship between two parties: therapist and patient. Successful psychotherapy depends on the positive progress of the correlation between these parties. This is because, most of the patients, usually involving children and adolescents, face problems with traditional psychotherapy [13]. It is difficult for young patients to develop an emotional connection with their therapist. Thus, it increases their resistance to sharing with the therapist. In turn, this difficulty could lead to unsuccessful treatment [14].

Utilising therapeutic games could provide rich experiences and is also capable of stimulating the motivation and engagement of patients, which are important during the psychotherapy session [15]. As an interactive medium, therapeutic games promise a viable, engaging and cost-effective approach that may benefit in reducing the stigma of mental illness [16]. Several therapeutic games have been purposely designed to enhance patients' motivation in order to support changes in their daily behaviour towards improving their quality of life. Therefore, a therapeutic game that is designed accordingly might increase one's intrinsic motivation and reduce reactance [17].

Several researchers have also proven that therapeutic game has a good prospect of supporting a higher level of cognition, for example, self-esteem, problem-solving, decision-making, cognitive and emotional skills [8]. Utilising therapeutic games in psychotherapy has already been proven as an efficient tool in supporting young patients during a psychotherapy session, capable of bringing positive changes to their mental health [18]. Hence, therapeutic games have high potential for improving health outcomes [13]. Playing games is synonymous with young patients, in which they can easily get immersed in the game. Hence, a strong relationship between therapist and patient can be built through this valuable tool.

A new style of communication between therapists and their patients can also be designed through therapeutic games, which can decrease face-to-face therapist contact [19]. This could very well contribute to a successful psychotherapy session. Activities involved in the session must be able to capture the attention of young patients. Thus, while they play and immerse in the game, they would give full attention to the game and forget that they are, in fact, in a psychotherapy session [20], [21]. At a certain point, they might even begin to feel comfortable with the psychotherapy environment and have no fear to express their feelings indirectly. Therefore, it is important for therapists to provide a convenient and safe environment during the therapeutic process [14].

## II. RELATED WORK

In this section, the most related work in therapeutic games for mental illness and targeted to young people from year 2021 to 2022 were discussed and compared. Based on the reviewed previous studies, it have clearly shown that the use of gaming intervention for mental illness in young people can improve their quality of life and reduce the depression symptoms. A comparative analysis that involved the 10 recent works related to therapeutic games in depression was conducted as shown in Table I.

Based on the comparative study, it was found that the most therapeutic approach utilized in the therapeutic games is Cognitive Behavior Therapy (CBT). This approach already known as an effective treatment to treat various mental health problems [25]. Thus, this is the main reason why most of the studies are using the therapeutic approach. This approach is very suitable to use for various game genres or platform such as video games, mobile games, online games, and role playing games. In addition, CBT is an action-oriented therapy that makes it very suitable for therapeutic games.

TABLE I. COMPARATIVE ANALYSIS

Game	Type	Therapeutic Approach	Focus	Source
I- SPARX	Video Game	CBT	Psychoeducation	[22]
SPARX	RPG	CBT	Behaviour change and engagement	[23]
Pesky gNATs (Depression & Anxiety)	Computer Game	CBT	Emotional problems	[24]
REThink (Depression)	Online Game	CBT REBT	Increase resilience	[8]
Grow It! (Depression)	Mobile Game	CBT	Emotional dynamics	[25]
Horizon: Resilience	Mobile Game	CBT Positive Psychotherapy	Increase motivation, cognitive flexibility activation and positivity	[26]
Moving Stories	Mobile 3D Video Game	NA	Mental health literacy and stigma reduction	[27]
EmoTIC	Mobile Game	NA	Social-emotional programme	[28]
Merlynne	Role Playing Game	CBT	Peer-to-peer support	[29]
MT-Phoenix i	Mobile Game	CBT	Reducing depressive symptoms	[30]

Legend:

CBT: Cognitive Behaviour Therapy

REBT: Rational Emotive Behavioral Therapy framework

Although all the studies as tabled in Table I focus on various elements, it clearly can be seen that most of them focus on the reducing the depressive symptoms by providing skills and knowledge to the young people. This is because young people have lack of skills and knowledge to handle depression. They also have a negative stigma toward depression as mentioned by [27]. Therefore, it is important to provide relevant skills and essential information to the young people to help them handle the depressive symptoms.

Other than that, gamification elements in the therapeutic game is vital to enhance engagement among the adolescents. Although the number of therapeutic games are increasing but the use of gamification in therapeutic games still limited. This is supported by a study conducted by [29]. The use of gamification elements will be able to sustain engagement among the adolescents to complete therapeutic activities in the game. The most common problem in the current treatment is inability to engage young people to the treatment and this leads to the dropping out from the treatment. Thus, the gamification elements should be fully utilized to increase their engagement.

### III. COGNITIVE BEHAVIOUR THERAPY

A broad range of therapeutic approaches is available for depression, but Cognitive Behaviour Therapy (CBT) has been the most widely used and extensively researched among young patients [25]. Fig. 1 describes the main components of this model.

This therapeutic approach is an effective intervention for improving coping strategies among adolescents. The effectiveness of CBT in treating depression among young patients is well-known and has been acknowledged by most therapists and researchers. To date, CBT has emerged as the 'gold standard' therapy approach for depression or even most mental illnesses [31]. In recent years, there is growing evidence for the efficacy of CBT on depression [32]. As depicted in Fig. 1, the components in CBT consist of: i) thoughts (how one thinks); ii) emotions (how one feels); and iii) behaviour (how one acts), combined to modify how adolescents think and react so as to eliminate negative thoughts.

The main aim of CBT is to assist patients in recognising their pattern of negative thinking, evaluating their validity, and replacing these faulty patterns with a more positive thinking style [33]. In the procedure of CBT with adolescents, therapists will observe and start analysing the patterns of thoughts, feelings and behaviour exerted by adolescents during particular events. This is done because therapists in CBT attempt to modify the ways adolescents think and feel in a more positive manner, which will be reflected in the behaviour exhibited by the patients in certain situations. Most importantly, CBT therapists would try to eliminate automatic negative thoughts that always influence the adolescents in their reactions.

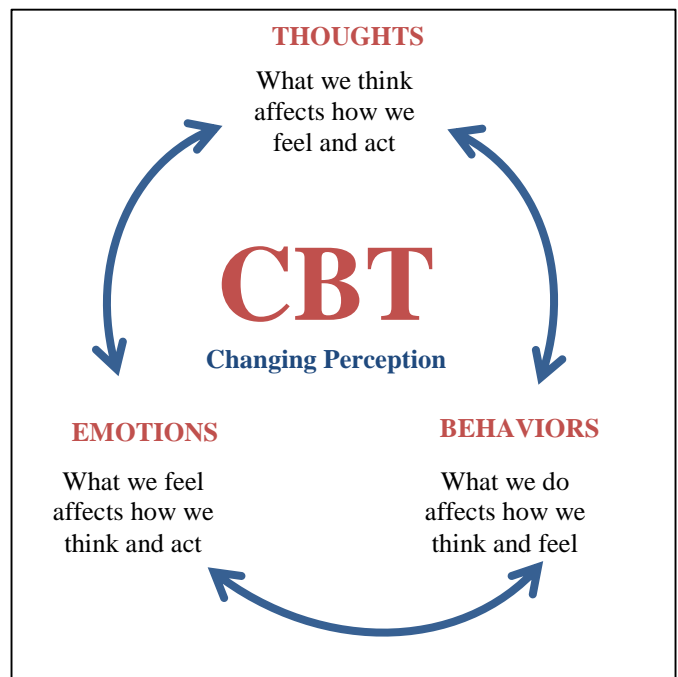


Fig. 1. Cognitive Behavior Therapy.

Now-a-days, CBT is widely utilised in therapeutic games that are designed for various mental illnesses, including depression [7]. This shows that the combination of CBT and therapeutic games can promote positive outcomes for people with depression, especially adolescents as tabled in Table I. To date, several therapeutic games were developed as an assistive aid to overcome the mental health illnesses. In this section, the study of the related works to therapeutic game discussed that focused on the reviewing the existing therapeutic game for mental illness.

### IV. CBT4DEPRESSION

The findings from the comparative analysis that was discussed in Section II lead to the design and development decision of CBT4Depression. CBT4Depression is an interactive therapeutic game application designed to deliver CBT to depressed adolescents. The target users for CBT4Depression were defined as ranging from 13 to 16 years old, who could play and are usually interested in computer games. It is also the most common age group among adolescents that is always experiencing depression. This therapeutic game application was meant to be a therapeutic material that serves as a kit for therapists or tutorials for the targeted users.

As found in the comparative analysis, gamification elements embed in the therapeutic game. Thus, the concept of CBT4Depression was designed to be fairly simple in a 2D game environment with a single character game control. The aim of CBT4Depression was to serve as an assistive aid to target users in identifying and reducing their depression levels. This is because vast evidence portrays that many adolescents, especially at present times, have failed to realise that they are suffering from depression [34].

Through this prototype application, users can learn techniques to strengthen their basic life skills that are critical to hinder their depression from getting worse. The CBT4Depression was designed and developed in the Malay language to suit the target users, as recommended by mental health experts during the preliminary investigation. So far, most therapeutic games have been developed in the English language, such as SPARX [19] and Pesky gNATS [24]. Table II summarises the detailed description of CBT4Depression.

### A. Therapeutic Elements

The CBT4Depression was designed based on a well-known therapeutic model named CBT, as described in Fig. 2. This model was chosen in this study because of its proven effectiveness in various mental health treatments. This model is also highly suitable for adolescents suffering from depression. CBT therapeutic strategies were applied to strengthen the elements of therapeutic game in CBT4Depression. The game highly focused on how to handle automatic negative thoughts and reactions.

TABLE II. DETAILED DESCRIPTION OF CBT4DEPRESSION

<b>Main topic</b>	Depression
<b>Target users</b>	Adolescents
<b>Type</b>	Single Player Role Playing Game (RPG)
<b>Game graphics</b>	2D Graphics
<b>Concept</b>	Therapeutic
<b>Language</b>	Malay
<b>Depression Inventory</b>	Beck Depression Malay
<b>Objective</b>	Beat Nega, collect points and escape from the jungle
<b>Therapy Content</b>	CBT, psychoeducation, and basic life skills

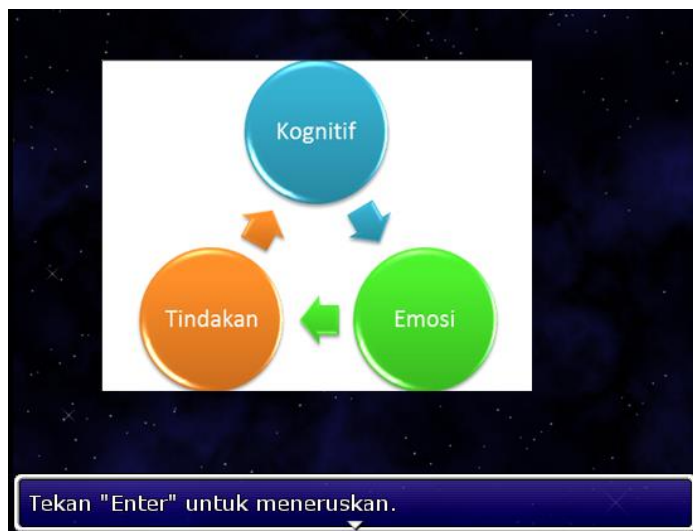


Fig. 2. CBT in CBT4Depression.

CBT4Depression helped the player recognise negative thoughts and poor reactions due to those thoughts. Thus, the therapeutic game embedded cognitive and behavioural techniques that are in line with CBT. Both techniques were implemented as part of the game elements and therapy content. Fig. 3 shows one of the cognitive techniques used in CBT4Depression, which is known as Socratic questioning.

Socratic questioning is an important component of CBT interventions that facilitates patients in assessing their automatic thoughts. It works by asking the player questions that encourage active participation in seeking answers and indirectly stimulating their critical thinking. Through this technique, the therapist can help patients become aware of and modify the process involved in their difficulties, and also learn how they can re-evaluate their thoughts.

Meanwhile, Fig. 4 describes one of the behaviour techniques called behaviour experiment, which was embedded in the therapeutic game. This technique was conducted after the player had learnt about negative thoughts and was used to evaluate underlying beliefs and assumptions.

The experiments were executed through questions and answers, writing notes on a certain given situation as an example and making predictions. Through these experiments, the player was encouraged to enhance the memory of the positive experience and avoid negative thoughts. Most of the existing therapeutic games for mental health apply these techniques in their games because such techniques are essential in psychotherapy [19], [35].

Besides, the use of a depression inventory was essential to measure the therapeutic outcome. A depression inventory is a set of self-rated questions used for assessing an individual's overall health condition related to depression symptoms [36]. In this study, the Beck Depression Inventory in Malay version (BDI-Malay) was utilized in the CBT4Depression. The BDI-Malay is suitable for adolescents because it is easy to understand. One example of the questions in the BDI-Malay is indicated in Fig. 5.



Fig. 3. Socratic Questioning.





Fig. 4. Behaviour Experiment.

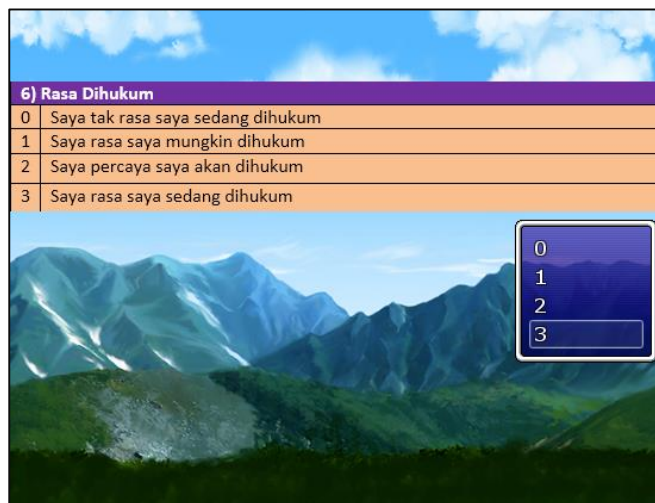


Fig. 5. BDI-Malay.

Other than that, psychoeducation was also embedded in the CBT4Depression to raise awareness among the target users. Concurrently, teaching critical basic life support skills in the therapeutic game application is also important in order to help target users cope with depression in certain situations. The life skills embedded in the CBT4Depression were problem-solving skill, decision-making skill, coping skill, relaxation skill, as well as communication and social skills. Similarly, existing therapeutic games that are designed for depressed adolescents also focus on comparable skills that cover cognitive, social and emotional skills for life and the school environment. The skills range from relaxation skills and literacy skills to emotion regulation skills [37].

### B. Game Elements

The storyline embedded in CBT4Depression was about a teenager (player) who was lost while exploring the jungle. The player needed to find a small village far inside the jungle to seek help from the villagers to escape from the jungle. However, at the same time, the player needed to fight with some enemies, called Nega. Nega(s) were incarnations from

the player's own negative thoughts that appeared in various forms, such as a friend, stepmother or black shadow. Hence, the player had to defeat Nega by fighting negative thoughts and modifying them into positive thoughts.

Various elements of mixed fantasy and curiosity were used in the game story to ensure that CBT4Depression was capable of capturing the interest and attention of the player as depicted in Fig. 6. The story involved the elements of exploration or fantasy, which then lead to surprise, wonder, and awe, all of these supporting the fun elements.

In addition, the rules in CBT4Depression were clear and specific, thus, allowing a player to receive feedback discrepancies, which then can trigger greater focused attention and enhance player engagement. The linking activities provided in the prototype also contributed to engaging competitive and cooperative motivations. All these elements can help in enhancing the engagement level among adolescents.



Fig. 6. Game Elements.

The difficulty levels in CBT4Depression were designed to increase gradually from the first level onwards. The difficulty at each level was matched with the player's skills so that the player can complete all the challenges in the game. The challenges in the therapeutic game were also tested by the health experts during evaluation to ensure that the moderate challenges were suitable for the target players [38]. As the therapeutic game was targeted at adolescents with depression, the challenges were not too difficult and yet not too easy.

Besides, since CBT4Depression was purposely designed for people with depression, time challenge was not utilised in the game based on the advice of the experts. After all, failing to meet the challenge may cause the player to feel more depressed. The challenges at each level required the player to master different skills. The relevant skills used by the player can be practised outside the game context, facilitating an effective skill transfer in the player's daily life [39]. Once the player mastered the skills at a level, the difficulty of the challenge will be increased at the next level, and the player will then need to master a new skill to complete the challenge. The challenges at each level are listed in Table III.

TABLE III. GAME CHALLENGES IN CBT4DEPRESSION

<b>Level 1</b>	Exploration, survival, and accuracy
<b>Level 2</b>	Exploration, memorisation, defeating the enemies, and finding a key to open an exit door.
<b>Level 3</b>	Defeating the enemy and helping a friend
<b>Level 4</b>	Answering questions correctly, finding essential information, and defeating the enemies.

Meanwhile, Fig. 7 shows one of the game challenges that was designed in the therapeutic game. In this game challenge, the player will learn that negative thinking will lower their confidence level and indirectly lead them in making poor decisions. Therefore, this game challenge could help to increase their awareness of the importance to hinder negative thinking.



Fig. 7. Game Challenges in Level 1.

### V. RESULTS AND DISCUSSION

In order to evaluate the effectiveness of CBT4Depression, a quasi-experimental study was conducted. The evaluation was adopted in this study to measure player experience in terms of engagement, as well as to test the effectiveness of CBT4Depression. The study was conducted in a lab environment with each PC installed with the CBT4Depression. Each participant in this study was assigned one PC.

A total of 115 adolescents aged between 13 and 16 were recruited in this study to measure their engagement level using CBT4Depression. The scores of BDI-Malay in the CBT4Depression were recorded and compared against the score classification, as indicated in Table III. The results demonstrated that the depression level of the respondents decreased from the pre to post-sessions. Then, the mean of the BDI score before and after using CBT4Depression was compared. Descriptive statistics for the two related samples were analysed, as presented in Table IV. The results showed that the mean for the BDI score after using the CBT4Depression was lower than the pre-score.

Nevertheless, the Paired Samples Test table was examined to ascertain whether the obtained result was significant or due to chance. Thus, the differences between the BDI score in both samples were examined for significance. As shown in Table V, the p-value was less than .05 (significance [2-tailed]). Referring to the results of the tests as shown in Table V and Table VI, it was found that the BDI scores reduced significantly after using CBT4Depression. Hence, this study has proven that therapeutic games support treatment for young people suffering from mental health, which is congruent to the findings of the study conducted by [13]. Therapeutic games can act as an alternative tools for psychotherapist in treating young patients such as children and adolescents. The information can be effectively delivered to them by using the therapeutic game as the medium of delivery.

TABLE IV. BDI SCORES CLASSIFICATIONS

Classification	Total Score	Pre-score	Post-score
Minimal Depression	0 - 9	-	79
Mild Depression	10 – 16	80	18
Borderline Clinical Depression	17 – 29	20	11
Moderate Depression	21 – 30	12	6
Severe Depression	31 – 39	3	1
Extremely Severe Depression	Over 40	-	-
<b>Total</b>		<b>115</b>	<b>115</b>

TABLE V. DESCRIPTIVE STATISTICS OF THE PAIRED SAMPLE (BDI SCORES)

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	Pre score	12.9739	115	8.22647	.76712
	Post score	8.0783	115	8.41901	.78508

TABLE VI. PAIRED SAMPLES TEST RESULT USING CBT4DEPRESSION

		Paired Differences				t	df	Sig. (2-tailed)	
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
					Lower				Upper
Pair 1	Pre score – Post score	4.89565	6.84717	.63850	3.63078	6.16052	7.667	114	.000

In addition, it is evident that the therapeutic elements and game elements in CBT4Depression is effective in reducing depression level among adolescents. In other words, CBT4Depression can be used as an assistive tool in mental health treatment among young patients.

## VI. CONCLUSION

This study has presented the effectiveness of CBT4Depression as a game-based digital intervention in reducing depression levels among adolescents. CBT4Depression comprises therapeutic elements that are adopted from a well-known therapeutic approach known as Cognitive Behaviour Therapy (CBT). In order to engage the adolescents in the game world, several game elements are applied in the game, such as challenge, curiosity, fantasy and fun. A quasi-experimental study has been conducted to measure the effectiveness of CBT4Depression. It has been found that CBT4Depression is effective in helping adolescents reduce their depression. Future work will involve expanding the CBT4Depression to cater more mental illness such as anxiety disorder, eating disorder and intermittent explosive disorders. The target groups will also expand to the adult as well.

## ACKNOWLEDGMENT

This study was supported by the Universiti Utara Malaysia through Journal Publication Fee Funding Scheme (SPYPJ).

## REFERENCES

- [1] World Health Organization, "Adolescent Mental Health," 2020. [Online]. Available: <https://www.who.int/news-room/factsheets/detail/adolescent-mental-health>. [Accessed: 13-Jun-2021].
- [2] A. B. Hink, X. Killings, A. Bhatt, L. E. Ridings, and A. Lintzenich, "Adolescent Suicide — Understanding Unique Risks and Opportunities for Trauma Centers to Recognize , Intervene , and Prevent a Leading Cause of Death," *Curr. Trauma Reports*, vol. 8, pp. 41–53, 2022.
- [3] R. Appleton, J. Gauly, F. Mughal, S. P. Singh, and H. Tuomainen, "Perspectives of young people who access support for mental health in primary care : A Systematic Review for Mental Health in Primary Care," *Br. J. Gen. Pract.*, vol. 72, no. 716, pp. E161–E167, 2022.
- [4] J. Goodwin, E. Savage, and A. O. Donovan, "I Personally Wouldn't Know Where to Go ": Adolescents' Perceptions of Mental Health Services," *J. Adolesc. Res.*, pp. 1–29, 2022.
- [5] S. O. Keffe, P. Martin, M. Target, N. Midgley, and G. A. Melvin, "I Just Stopped Going ": A Mixed Methods Investigation Into Types of Therapy Dropout in Adolescents With Depression," *Front. Psychol.*, vol. 10, no. 75, pp. 1–14, 2019.
- [6] N. I. Othman, N. A. M. Zin, and H. Mohamed, "Play-centric designing of a serious game prototype for low vision children," *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 5, pp. 199–205, 2020.
- [7] A. Dewhirst, R. Laugharne, and R. Shankar, "Therapeutic use of serious games in mental health : scoping review," *BJPsych Open*, vol. 8, no. 2, p. A238, 2022.
- [8] O. A. David, S. Magurean, and C. Tomoiaga, "Do Improvements in Therapeutic Game-Based Skills Transfer to Real Life Improvements in Children' s Emotion-Regulation Abilities and Mental Health? A Pilot Study That Offers Preliminary Validity of the RETHink In-game Performance Scoring," *Front. Psychiatry*, vol. 13, no. March, p. 828481, 2022.
- [9] M. Kowal, E. Conroy, N. Ramsbottom, and T. Smithies, "Gaming Your Mental Health: A Narrative Review on Mitigating Symptoms of Depression and Anxiety Using Commercial Video," *JMIR Serious Games*, vol. 9, no. 2, p. e26575, 2021.
- [10] J. Torous et al., "The growing field of digital psychiatry: current evidence and the future of apps, social media, chatbots, and virtual reality," *World Psychiatry*, vol. 20, no. 3, pp. 318–335, 2021.
- [11] M. Fitzgerald, B. Sc, M. Sc, G. Ratcliffe, B. Sc, and M. Sc, "Serious Games, Gamification, and Serious Mental Illness: A Scoping Review," *Psychiatr. Serv. Adv.*, vol. 71, no. 2, pp. 170–183, 2020.
- [12] M. Fitzgerald, B. Sc, M. Sc, G. Ratcliffe, B. Sc, and M. Sc, "Serious Games , Gami fi cation , and Serious Mental Illness: A Scoping Review," no. February, 2020.
- [13] H. Van Der Meulen, D. Mccashin, G. O. Reilly, and D. Coyle, "Using Computer Games to Support Mental Health Interventions : Naturalistic Deployment Study," *JMIR Ment. Heal.*, vol. 6, no. 5, p. e12430, 2019.
- [14] E. Wilmots, N. Midgley, L. Thackeray, S. Reynolds, and M. Loades, "The therapeutic relationship in Cognitive Behaviour Therapy with depressed adolescents : A qualitative study of good-outcome cases," *Psychol. Psychother. Theory, Res. Pract.*, 2019.
- [15] H. W. Wong et al., "Postsecondary student engagement with a mental health app and online platform (Thought spot): Qualitative study of user experience," *JMIR Ment. Heal.*, vol. 8, no. 4, pp. 1–12, 2021.
- [16] R. R. Wehbe et al., "Designing a Serious Game (Above Water) for Stigma Reduction Surrounding Mental Health: Semistructured Interview Study With Expert Participants," *JMIR Serious Games*, vol. 10, no. 2, p. e21376, 2022.
- [17] A. Fuchslocher, J. Niesenhaus, and N. Krämer, "Serious games for health: An empirical study of the game 'Balance' for teenagers with diabetes mellitus," *Entertain. Comput.*, vol. 2, no. 2, pp. 97–101, Jan. 2011.
- [18] D. Zayeni, J. Raynaud, and A. Revet, "Therapeutic and Preventive Use of Video Games in Child and Adolescent Psychiatry : A Systematic Review," *Front. Psychiatry*, vol. 11, no. 36, 2020.
- [19] T. M. Fleming et al., "SPARX-R computerized therapy among adolescents in youth offenders' program : Step-wise cohort study," *Internet Interv.*, vol. 18, no. September, p. 100287, 2019.
- [20] S. Henrich and R. Worthington, "Let Your Clients Fight Dragons: A Rapid Evidence Assessment regarding the Therapeutic Utility of 'Dungeons & Dragons,'" *J. Creat. Ment. Heal.*, vol. 00, no. 00, pp. 1–19, 2021.
- [21] J. Steadman, C. Boska, C. Lee, X. S. Lim, and N. Nichols, "Using Popular Commercial Video Games in Therapy with Children and Adolescents," *J. Technol. Hum. Serv.*, vol. 32, no. 3, pp. 201–219, 2014.
- [22] A. Thomas, Y. Bohr, J. Hankey, M. Oskaln, J. Barnhardt, and C. Singoorie, "How did Nunavummiut youth cope during the COVID-19 pandemic? A qualitative exploration of the resilience of Inuit youth leaders involved in the I-SPARX project," *Int. J. Circumpolar Health*, vol. 81, no. 1, 2022.
- [23] T. Fleming, M. Lucassen, K. Stasiak, K. Sutcliffe, and S. Merry, "Technology Matters: SPARX – computerised cognitive behavioural therapy for adolescent depression in a game format," *Child Adolesc. Ment. Health*, vol. 26, no. 1, pp. 92–94, 2021.
- [24] D. McCashin, D. Coyle, and G. O'Reilly, "Pesky gNATs for children experiencing low mood and anxiety – A pragmatic randomised controlled trial of technology-assisted CBT in primary care," *Internet Interv.*, vol. 27, no. December 2021, p. 100489, 2022.
- [25] E. Dietvorst, M. A. Aukes, J. S. Legerstee, A. Vreeker, and M. Micah, "A Smartphone Serious Game for Adolescents (Grow It! App): Development, Feasibility, and Acceptance Study," *JMIR Form. Res.*, vol. 6, no. 3, p. e29832, 2022.
- [26] Á. Gómez-cambronero, S. Casteleyn, and A. Mira, "Horizon : Resilience – Design of a Serious Game for Ecological Momentary Intervention for Depression," in *Extended Abstracts of the 2021 Annual Symposium on Computer-Human Interaction in Play (CHI PLAY '21)*, 2021, pp. 236–241.
- [27] A. Tuijnman, M. Kleinjan, M. Olthof, E. Hoogendoorn, I. Granic, and R. C. Engels, "A Game-Based School Program for Mental Health Literacy and Stigma on Depression (Moving Stories): Cluster Randomized Controlled Trial," *JMIR Ment. Heal.*, vol. 9, no. 8, p. e26615, 2022.
- [28] U. de la Barrera, E. Mónaco, S. Postigo-Zegarra, J. A. Gil-Gómez, and I. Montoya-Castilla, "EmoTIC: Impact of a game-based social-emotional



- programme on adolescents,” PLoS One, vol. 16, no. 4 April, pp. 1–17, 2021.
- [29] T. Chan, R. P. Gauthier, A. Suarez, N. F. Sia, and J. R. Wallace, “Merlynn: Motivating Peer-to-Peer Cognitive Behavioral Therapy with a Serious Game,” Proc. ACM Human-Computer Interact., vol. 5, no. CHIPLAY, p. Article 250, 2021.
- [30] G. Costikyan, “I Have No Words & I Must Design: Toward a Critical Vocabulary for Games,” in Proceedings of Computer Games and Digital Cultures Conference, 2002, pp. 9–33.
- [31] P. Srivastava, M. Mehta, R. Sagar, and A. Ambekar, “smartteen- a Computer Assisted Cognitive Behavior Therapy for Indian Adolescents with Depression- A Pilot Study,” Asian J. Psychiatr., p. 101970, 2020.
- [32] H. D. Hadjistavropoulos et al., “An Internet-Delivered Cognitive Behavioral Therapy for Depression and Anxiety Among Clients Referred and Funded by Insurance Companies Compared With Those Who Are Publicly Funded: Longitudinal Observational Study,” JMIR Ment. Heal., vol. 7, no. 2, p. e16005, 2020.
- [33] S. A. Bhat, “Cognitive Behavioral Therapy and Depression,” Int. J. Adv. Educ. Res., vol. 2, no. 6, pp. 143–145, 2017.
- [34] E. Eigenhuis et al., “Facilitating factors and barriers in help-seeking behaviour in adolescents and young adults with depressive symptoms: A qualitative study,” PLoS One, vol. 16, no. 3 March 2021, pp. 1–20, 2021.
- [35] H. van der Meulen, G. O’Reilly, and D. Coyle, “Including End-Users in Evaluating and Designing a Game that Supports Child Mental Health,” in Proceedings of the 2018 Annual Symposium on Computer-Human Interaction in Play Companion Extended Abstracts - CHI PLAY ’18 Extended Abstracts, 2018, pp. 655–659.
- [36] M. Okada, Y. Nakadoi, and A. Fujikawa, “Relationship between self-rated health and depression risk among children in Japan,” Humanit. Soc. Sci. Commun., vol. 9, no. 1, pp. 1–9, 2022.
- [37] T. Kuosmanen, T. M. Fleming, J. Newell, and M. M. Barry, “A pilot evaluation of the SPARX-R gaming intervention for preventing depression and improving wellbeing among adolescents in alternative education,” Internet Interv., vol. 8, pp. 40–47, Jun. 2017.
- [38] P. M. de Souza, K. R. da Hora Rodrigues, F. E. Garcia, and V. P. de Almeida Neris, “Towards a Semiotic-Based Approach to the Design of Therapeutic Digital Games,” in Digitalisation, Innovation, and Transformation. ICISO 2018. IFIP Advances in Information and Communication Technology, vol. 527, K. Liu, K. Nakata, W. Li, and C. Baranauskas, Eds. Springer International Publishing, 2018, pp. 53–62.
- [39] M. Poppelaars, A. Lichtwarck-Aschoff, M. Kleinjan, and I. Granic, “The impact of explicit mental health messages in video games on players’ motivation and affect,” Comput. Human Behav., vol. 83, no. 2018, pp. 16–23, 2018.

# Creating Video Visual Storyboard with Static Video Summarization using Fractional Energy of Orthogonal Transforms

Ashvini Tonge<sup>1</sup>

Department of Information Technology  
Pimpri Chinchwad College of Engineering  
Pune, India

Sudeep D. Thepade<sup>2</sup>

Department of Computer Engineering  
Pimpri Chinchwad College of Engineering  
Pune, India

**Abstract**—The overwhelming number of video uploads and downloads has made it incredibly difficult to find, gather, and archive videos. A static video summarization technique highlights an original video's significant points through a set of static keyframes as a video visual storyboard. The video visual storyboards are created as static video summaries that solve video processing-related issues like storage and retrieval. In this paper, a strategy for effectively summarizing static videos using the feature vectors, which are fractional coefficients of the transformed video frames, is proposed and evaluated. Four popular orthogonal transforms are deployed for generating feature vectors of video frames. The fractional coefficients of transformed video frames taken as 25 percent, 6.25 percent, and 1.5625 percent of full 100 percent transformed coefficients are considered to form video visual storyboards. The proposed method uses the benchmark video datasets Open Video Project (OVP) and SumMe to validate the performance, containing user summaries (storyboards). These video summaries created using the proposed method are evaluated using percentage accuracy and matching rate.

**Keywords**—Keyframe; orthogonal transform; VSUMM; video visual storyboard; video summarization

## I. INTRODUCTION

Due to the significant increase in the cases of pushing promotional videos in online drop boxes, Emails, and social network accounts of users, the users are forced to get these videos downloaded to understand the contents of the video. After seeing the video, the user often finds that he is not interested in the whole content. With the easy accessibility of Internet Services and handheld image/video capturing devices, there is a lot increase in the photos and videos in online/offline databases. But it has introduced new challenges in computer vision research, such as storage, search, and navigation, due to the huge volume of video data. There is a critical need to address these problems because of the abundance and accessibility of video data. A video content summarization aims to summarize the full video content in this situation into short video clips or groups of frames that are crucial for understanding video content. This summary is known as a visual video storyboard.

Video content summarization through storyboards enables quick browsing of a collection of sizable video datasets. Additionally, it supports associated video-related tasks like

video indexing and retrieval. Nowadays, video summarization has evolved in various applications as a problem of keyframe extraction [1]. But the key frame extraction is very challenging due to the complex nature of the video. Key frame extraction, which substitutes for the most crucial elements of the movie, is one method for producing video summaries/storyboards.

The video storyboard is a quick and meaningful way of giving an abstract perspective on an entire video by creating a video summary[2]. The viewer might not have enough time to see the complete movie. At that time, the storyboards may help users to watch only the important content using these keyframes to narrate the full story of a video. These storyboards can be static [3] or dynamic [4].

Over the past two to three decades, videos have increased. Still, there isn't an ideal system that can handle the time-consuming process of creating a visual video storyboard. So, the indexing, retrieval, and storage of video are affected. All of these video-related concerns can be addressed by video storyboards. The number of approaches proposed for creating video summaries mainly focuses on feature selection techniques used for keyframe selection and evaluation with the ground truth.

The existing video summarization methods divide the whole video into shots and segments. Then it applies the feature selection process as defined in VSUMM [5][6][7], the DT triangulation method for clustering the video frames [8], Local descriptor based and temporal features based [9], diverse color space-based key frame extraction [10].

In VSUMM and DT methods, the video frames are considered in a batch of the first 25-30 frames or by interleaving sequence, thereby not including all the features in video frames. In VSUMM and DT methods, the keyframes are selected only by grouping and distance between the video frames. This leads to the loss of a few important frames in a sequence. This limitation can be overcome by using all the video frames in a video, as stated in the proposed method.

In state-of-art summarization of video, video summaries are dependent on the key frame extraction, and feature selection plays an essential role in this keyframe extraction step. Therefore many researchers have demonstrated different techniques for selecting these features for key frame extraction

[11][12]. Based on the input and output features, video summarization has two different ways: dynamic and static. A dynamic summary of the video is the abstraction of the lengthy video into a compact reel in which the scene is recreated using only keyframes, and a motion is applied [13]. A static summary of the video includes a series of static keyframes that shows the entire story of the video without motion. The choice of static and dynamic is user dependent.

The use of orthogonal transforms assures the full feature in the input frames. Orthogonal transforms are applied with fractional energy coefficients for the various applications of content-based video retrieval with performance measure as precision and recall[14]. The use of transformed features assures high energy compaction; therefore, transformed features are used in video processing. The research work presented here addresses the issue of feature selection in key frame extraction by using the transformed features and proposes a novel video summarization technique with the creation of visual video storyboards. The proposed method first segments the video into video frames, spreading all video content over multiple frames.

In most of the existing static video summarization approaches, the observed limitations are the huge size of feature vectors of video frames, unequal size of feature vectors, suitability of the features for a particular type of video dataset only, and experimental validation is done with a single dataset. Hence there is a need to have the optimal minimum size of a more robust feature vector with the ability to show analogous performance across multiple video datasets.

Depending upon the discussion above, the key contributory significance of the proposed static video summarization method is as follows:

- 1) The use of fractional energy of transformed video frames to produce a video summary (video visual storyboard).
- 2) The use of orthogonal transforms to obtain the fractional coefficients of transformed video frames.
- 3) Performance validation using Open Video Project (OVP) and SumMe benchmark video datasets.

The remaining part of this paper is structured as follows: Section 2 illustrates the current work. The proposed static video summarization method using fractional energy coefficients of transformed video frames is put forth in Section 3, and Section 4 describes the results with the OVP and SumMe dataset and the test bed used for experimentation. The conclusion obtained by thorough investigation and demonstration is summarised in Section 5.

## II. RELATED WORK ON STATIC VIDEO SUMMARIZATION

Lengthy videos have a large sequence of short segments (shots) in video frames; these shots are made up of only the most essential frames (keyframes) that can be used to search and retrieve the original videos. Through these keyframes (storyboards), the videos can be understood easily. According to the literature, transformed features ensure retrieval effectiveness and reduce the calculations required for time-consuming video processing [14]. But these transformed features are extracted for content-based image retrieval, not video retrieval. It does not include a sequence of images in a

query. In [15], adaptive threshold-based key frame extraction uses the MPEG -7 color layout descriptors combined with adaptive thresholding. In [16], annotation-based keyframe identification is defined with interest as a key frame identification concept. Both static video frame extraction methods use the actions in a video as a base for the shot selection. This will not apply to all videos; a few videos may be just informative or storytelling. In [17], a review presents different video summarization categories based on features, clusters, shots, and trajectories. But this study concludes with a video summarization of the region of interest problem. Every time human intervention is needed while summarizing the video. So there is a need for an automatic summary generator with minimum computations. The specific feature should be selected with automatic computations for the keyframe selection.

Orthogonal transforms, including Discrete Cosine, Kekre, Walsh, Slant, Discrete Sine, and Discrete Hartley, have been explored for content-based image retrieval (CBIR)[18]. Mean Square Error (MSE) is the similarity metric considered. This method creates an efficient image signature for each image and ensures full input feature selection. But this operation is performed on each distinct image in a dataset. In a video, many similar images, known as near duplicates, will increase the computational overhead; in such cases, using transformed features may reduce the computational complexity. Therefore in the proposed work, different orthogonal transforms are used for storyboard creation in static video content summarization.

But in [18], the transformed features are used for image retrieval; no recreation is performed here with the retrieved image. The proposed video visual storyboard creation method is explained in detail in the following section and generated video storyboards are validated using the novel performance metrics. A brief review of the techniques that support the video summarization is given in Table I.

TABLE I. RELATED WORK COMPARISON OF VIDEO SUMMARIZATION TECHNIQUES

Author List	Type of Features used	Dataset	Performance
Xiang et al. [16] 2020	ConvNet	VSUMM	F-score (72.1%)
Rukiye et al. [19] 2021	CNN & RNN	UCF 101	Accuracy (67.39%)
Vijay Kumar et al. [20] 2014	Discrete Wavelet Transform, Haar Wavelet based	Sports Video	Precision (0.83)
Naveed et.al. [21] 2013	Discrete Cosine Transform	Open Video Project	F-Measure (82%)
Kavitha et al. [22] 2015	Discrete Wavelet Transform	Open Video Project	F1-Score (87%)
Ajay Narvekar et al. [23] 2013	Discrete Cosine Transform	Online videos	Precision (0.78)

The work presented in [19][20][21][22], briefly compared in Table I, clearly indicates that the transformed feature gives more precision and recall than the cluster-based approach,

along with the reduction in computational costs since kernel size varies. It provides a quick review of existing methods where the orthogonal transforms are used in video summarization, and CNN performance is also compared with recent work. This has given the motivation for selecting the orthogonal transforms to prove the efficiency for static video content summarization, i.e., creating the video visual storyboards.

### III. PROPOSED METHOD OF STATIC VIDEO SUMMARIZATION USING ORTHOGONAL TRANSFORMS

The proposed method of creating feature vectors for static video summarization uses fractional energy coefficients of transformed features to make a video visual storyboard. The proposed framework is shown in Fig. 1.

The Proposed framework has three steps: first, to form a transformed feature vector of all video frames using the 'T' transform; the second step is to prepare a feature vector using fractional energy coefficients; the third step is to select the number of keyframes. All these three steps are pictorially represented in Fig. 1. The above steps are elaborated in the following subsections.

#### A. Orthogonal Transform

Different orthogonal transforms used in this proposed method are discussed here. The 'T' transforms used in this system are defined in the form of their matrix equations.

##### 1) Discrete Cosine Transform (DCT)

The Discrete Cosine Transform is the most widely used orthogonal transform in image processing. The  $N \times N$  cosine transform matrix is defined as below in equation (1),

$$c(p,n) = \begin{cases} \frac{1}{\sqrt{N}} & p = 0, 0 \leq n \leq N-1 \\ \sqrt{\frac{2}{N}} \cos \frac{\pi(2n+1)p}{2N} & 1 \leq p \leq N-1, 0 \leq n \leq N-1 \end{cases} \quad (1)$$

##### 2) Slant Transform

The Slant transform is a constant function with a one-row function and the second row is a linear function of the column index. It includes sparse matrices, reducing the computations and leading to a fast process.

The matrix equation of the Slant transform is given by equation (2),

$$S_2 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix} \quad (2)$$

##### 3) Walsh Transform

The set of  $N$  rows denoted as  $W_k$ , for  $k=0, 1, \dots, N-1$ , is defined in Walsh Matrix that has distinct properties.

$W_k$  takes on the values +1 or -1.

$W_k [0] = 1$  for all  $k$ .

$W_k \times W_l = 0$ ,  $k \neq l$  and  $W_k \times W_l$  has exactly  $k$  zero crossings, for  $k=0, 1, \dots, N-1$ .

Each  $W_k$  is either even or Odd.

##### 4) Kekre Transform

This transform matrix is an  $N \times N$  matrix, where the upper diagonal values are one, and the diagonal values of Kekre's transform matrix are also one, except other values below the diagonal is zero.

This matrix equation is defined using the Hadamard matrix of order  $N$  in equation (3),

$$K_{xy} = \begin{cases} 1 & , x \leq y \\ -N + (x+1) & , x = y+1 \\ 0 & , x > y+1 \end{cases} \quad (3)$$

#### B. Feature Vector Extraction

Each video frame is resized to  $256 \times 256$ . On each color plane video frame of size  $N \times N$ , the 'T' Transform (alias DCT, Walsh, Slant, and Kekre Transform) is applied to extract the visual feature vector of size  $N \times N$  as a full or 100% energy content scenario as shown in Fig. 1.

The fractional energy coefficients are computed by dividing the full features of the video frame into block sizes of  $32 \times 32$ ,  $64 \times 64$ , and  $128 \times 128$  and are taken as top left-hand side coefficients of transformed color planes of video frames, as shown in Fig. 2.

In this proposed method, the transformed coefficients are used to form the feature vector. In [24], transformed coefficients as features have shown better accuracy in the keyframe extraction. The proposed method uses these transformed video frame coefficients with a reduced number of feature vector elements.

#### C. Feature Vector Database using Fractional Coefficients

The proposed video visual storyboard generation method for static video summarization uses fractional energy coefficients. The diagrammatic representation of extracting fractional energy coefficients to generate feature vectors from a transformed video frame is shown in Fig. 2.

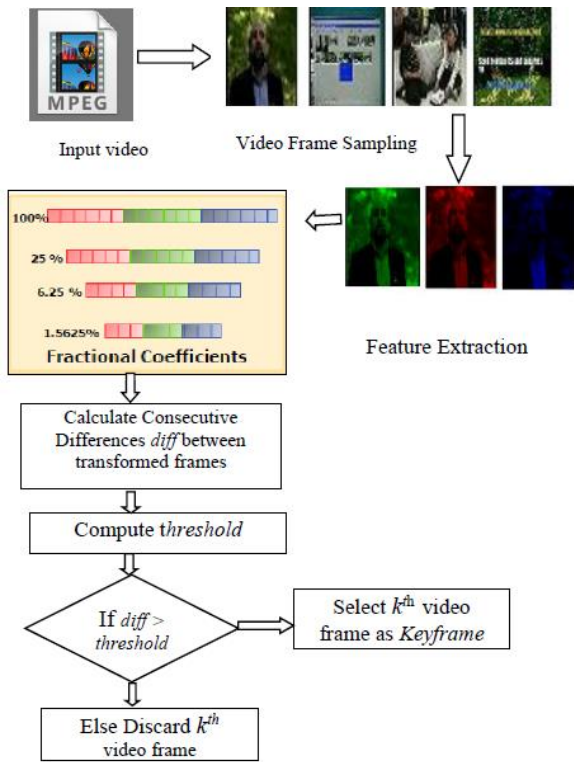


Fig. 1. The Proposed Method of KeyFrame Extraction with Fractional Energy Coefficients of Transformed Video Frames for Visual Storyboard Creation.

If the video frame is of size 256x256, the fractional energy coefficient proportions are taken as 25%, 6.125%, and 1.5625%, respectively, with sizes 128x128, 64x64, and 32x32. Considering high energy coefficients as feature vectors reduces feature vector size, time, and computational complexity.

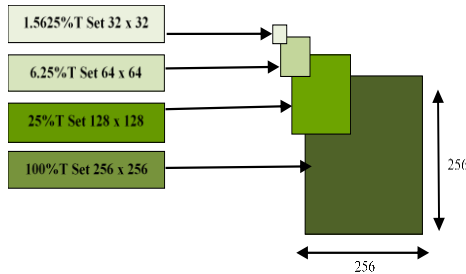


Fig. 2. Proposed Feature Vector Extraction with Fractional Energy Coefficient.

#### D. The Decision of Keyframe for Video Visual Storyboard

The keyframes are the significant frames that contain the maximum information in a video frame. These keyframes are selected based on the consecutive differences between two transformed video frames. These consecutive differences are then compared with the certain constant threshold calculated with standard deviation and mean. Each transformed video frame in the sequence is associated with the difference; if this difference is above the threshold, then that particular frame is selected as a keyframe; otherwise, it is discarded. In this process of keyframe selection, there might be a possibility of selecting near duplicates along with essential frames. These

near duplicates are eliminated by interleaving manually or statistically.

TABLE II. VIDEO DATASET DETAILS CATEGORY WISE

Open Video Project Video Dataset			
Category	Documentary	Educational	Lecture
*No. of videos	15	20	20
Category	Historical	Public Service	Ephemeral
*No. of videos	20	20	15
Total 110 videos			
SumMe Video dataset			
Total 25 videos			

#### IV. RESULTS AND DISCUSSION

The implementation details and results are discussed in this section.

##### A. Experimental Video Testbed

Two video datasets, the Open Video Project (110 videos) and SumMe (25 videos) of various categories, are used for proposed experimentation. The videos are provided in both compatible file types (MPEG-2, MPEG-4). These video datasets are openly accessible. The OVP contains the categories of Lectures, Television, Demonstrations, and Documentary videos. The SumMe dataset includes videos of categories like Cooking, Bike polo, Base jumping, etc.

These two benchmark datasets provide video user summaries as visual storyboards for respective videos [21]. These user summaries are further used for performance comparison to evaluate the proposed static video summarization technique for creating video visual storyboards. A few video frames are shown in the following Fig. 3, with a few sample frames from the OVP and SumMe datasets. Each video length varies from less than 1 minute to 2 minutes.



a) OVP- Family TV Spots around the World



b) SumMe – AirForce, Base jumping, fire Demo

Fig. 3. Few Video Frames from a) OVP Dataset and b) SumMe Dataset.

Table II shows the number of videos considered from the OVP and SumMe datasets for the experimentation of the proposed method with respective categories. The type of video in the Lecture category has slow transitions in the scene, leading to fewer keyframes in the final output of the video visual storyboard. The sudden changes in the scene will affect the number of keyframes extracted. The proposed experimentation testbed includes all types of videos with such variations.

### B. Results and Discussion

The experimentation is performed using the above videos shown in Fig. 3. The proposed method extracts the keyframes from each video from the dataset. The obtained keyframes are compared with the already existing given ground truth in the OVP and SumMe datasets. The performance metrics matching rate and percentage accuracy are calculated to identify the exact matching frame from the given set of keyframes in the OVP and SumMe video dataset storyboard.

The number of keyframes extracted using the fractional energy coefficients of transformed video frames is evaluated using the given ground truth of videos from the OVP and SumMe datasets.

#### 1) Performance Metric

The percentage accuracy is calculated as the ratio of the number of correctly extracted keyframes by the proposed method and the total number of keyframes given in the standard user storyboard.

In the matching rate, the matching from the given summaries with frame numbers given in ground truth is done with the keyframes obtained using the proposed method. The matching rate is calculated as the number of identical matching video frames similar to the keyframes given in the OVP video user summary. Here it is assumed that keyframes in the user summary are provided with the frame numbers from the original video frame sequence.

The performance metrics used in the proposed system are explained in equations (4) and equation (5).

The keyframes in the given OVP summary are downloaded from <https://openvideo.project.com>. The user summary from OVP and SumMe datasets are compared with a set of keyframes obtained through the proposed system.

The keyframes obtained are used to create a visual video storyboard. The results are summarized in Tables III and IV for the orthogonal transform using fractional energies with OVP and SumMe videos.

The percentage accuracy and matching rate are given in Tables III and IV with the detailed analysis of the proposed fractional energy-based keyframe extraction method using orthogonal transform alias DCT, Walsh, Slant, and Kekre transform.

The results show that the performance improves in the case of the proposed use of fractional energy coefficients compared to the consideration of 100% coefficients. This reduction in the sizes of the different feature vectors in the proposed method improves the accuracy of video visual storyboard creation by finding more accurate keyframes.

TABLE III. PERCENTAGE ACCURACY AND MATCHING RATE OF PROPOSED FRACTIONAL ENERGY COEFFICIENTS BASED ON KEYFRAME EXTRACTION METHOD FOR RESPECTIVE ORTHOGONAL TRANSFORMS EXPERIMENTED ON OVP DATASET

Performance using OVP dataset				
Fractional energy coefficients	10 0%	25 %	6.25 %	1.526 %
Discrete Cosine Transform (DCT)				
% Accuracy	76. 23	75. 31	75.0 6	76.28
Matching Rate	19. 2	20	20.2	20.2
Walsh Transform				
% Accuracy	73. 45	73. 49	72.2 5	73.63
Matching Rate	16. 63	17. 24	17.4 8	16.78
Slant Transform				
% Accuracy	70. 19	71. 59	70.3 6	71.45
Matching Rate	15. 82	15. 24	15.6 3	15.89
Kekre Transform				
% Accuracy	69. 38	70. 39	70.7 2	70.49
Matching Rate	14. 92	15. 24	15.7 8	15.58

The above performance comparison, as shown in Table III and Fig. 4 indicates that the results of DCT based proposed keyframe extraction method outperform when compared to the other Walsh, Slant, and Kekre orthogonal transform-based keyframe extraction.

$$\text{Percentage Accuracy} = \frac{\text{Number of correctly extracted keyframes}}{\text{Expected number of keyframes}} \quad (4)$$

$$\text{Matching Rate} = \frac{\text{Exact matching extracted keyframes}}{\text{Total number of keyframes}} \quad (5)$$



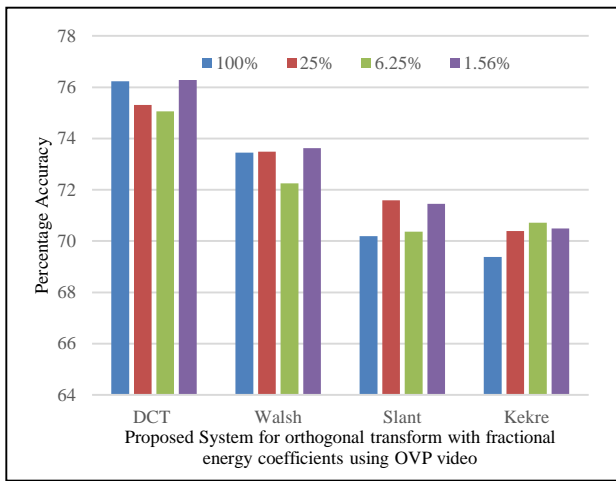


Fig. 4. Performance Comparison of the Proposed Fractional Energy Coefficients based Video Keyframe Extraction Method for Respective Orthogonal Transforms Experimented on OVP Dataset.

A similar method applies to other "T" orthogonal transforms with fractional energy coefficients. The reduction in the feature vector is not affecting the percentage accuracy but increases the selection of keyframe matching to the given keyframe from OVP visual storyboards. Table IV and Fig. 5 show the analysis of the performance obtained by the proposed method using SumMe videos.

TABLE IV. PERCENTAGE ACCURACY AND MATCHING RATE OF PROPOSED FRACTIONAL ENERGY COEFFICIENTS BASED ON KEYFRAME EXTRACTION METHOD FOR RESPECTIVE ORTHOGONAL TRANSFORMS EXPERIMENTED ON SUMME DATASET

Performance using SumMe dataset				
Fractional energy coefficients →	10	25	6.25	1.526
	0%	%	%	%
<b>Discrete Cosine Transform (DCT)</b>				
% Accuracy	.75	.74	.743	.7359
Matching Rate	.18	.17	.179	.175
<b>Walsh Transform</b>				
% Accuracy	.73	.73	.722	.7363
Matching Rate	.16	.17	.174	.1678
<b>Slant Transform</b>				
% Accuracy	.72	.72	.731	.7324
Matching Rate	.16	.15	.162	.1649
<b>Kekre Transform</b>				
% Accuracy	.73	.73	.742	.7347
Matching Rate	.16	.16	.154	.1638

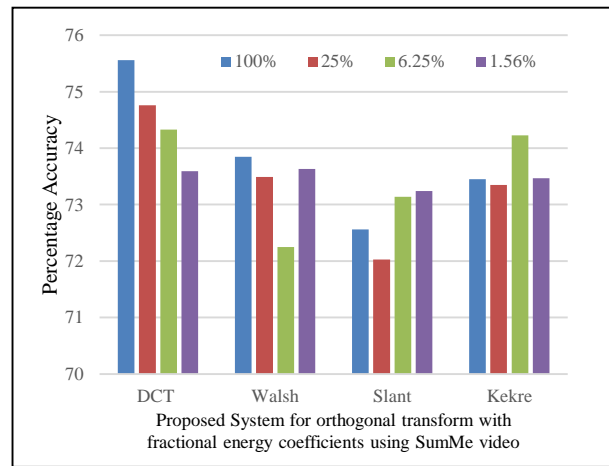


Fig. 5. Performance Comparison of the Proposed Fractional Energy Coefficients based Video Keyframe Extraction Method for Respective Orthogonal Transforms Experimented on SumMe Dataset.

### 2) Significance of the Proposed Method

The proposed method generates video summaries with a few keyframes displayed below in Fig. 6(a) and 6(b) to support the performance metrics discussed in this paper. The similarity can be compared with the frame numbers similar to the OVP and SumMe storyboards.

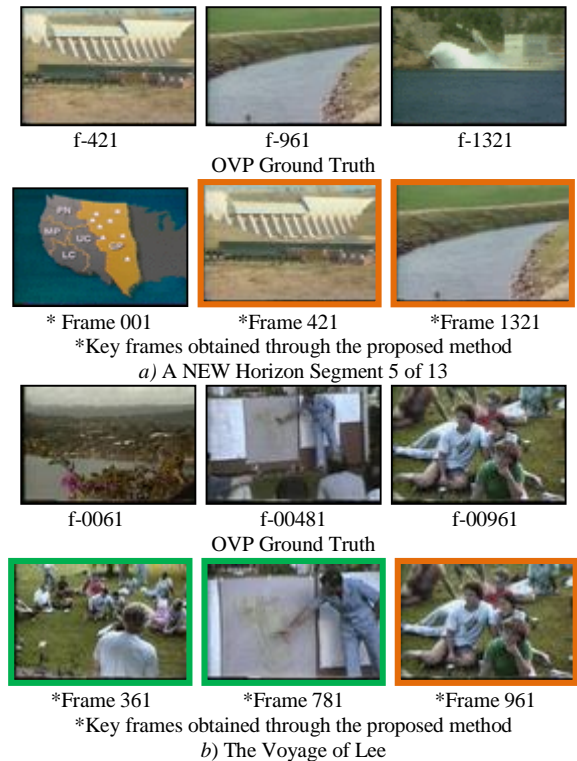


Fig. 6. Video Frames of a Video – a) A NEW Horizon Segment 5 of 13 b) The Voyage of Lee (Highlighted Keyframes Match with OVP Storyboard).

The highlighted frames, as shown in Fig. 6 above, are the ones that are used to calculate the matching rate. The keyframes whose frame numbers match the given keyframes in the OVP storyboard are highlighted.

The relationship between the feature vector computation and the energy reduction coefficients utilized in this implementation is shown in Table V. The 6.25% feature vector space reduces the whole feature vector by 93.75% and gives similar percentage accuracy.

TABLE V. COMPARISON OF PROPOSED % FRACTIONAL ENERGY COEFFICIENTS AND REDUCTION IN % FEATURE VECTOR SIZES

Feature Vector Size	% fractional Energy Coefficients	% reduction in Feature Vector Size
$N \times N \times 3$	100	0
$\frac{N}{2} \times \frac{N}{2} \times 3$	25	75
$\frac{N}{4} \times \frac{N}{4} \times 3$	6.25	93.75
$\frac{N}{8} \times \frac{N}{8} \times 3$	1.5625	98.4375

The proposed method here is effective for storyboard generation as compared to other techniques DT [8] and VSUMM [7] in terms of computations required to process video frames. Here the dimensionality of each feature vector is reduced due to the use of fractional energy coefficients.

1) Comparison with other Techniques

This section compares the proposed system with existing techniques like DT [8] and OVP summary. The experiments were performed on videos downloaded from OVP and SumMe. These summaries are evaluated in percentage accuracy and compared with existing VSUMM and DT Summary ground truth. The same comparison of the proposed method performance is made with VSUMM and DT summaries using the performance metric as percentage accuracy. Tables VI and VII show this comparison.

Fig. 7 below shows the visual static storyboard comparison between OVP ground truth, VSUMM, and DT summary with the static summary obtained through the proposed method of static video summarization in the form of a set of keyframes extraction for storyboard creation. Fig. 7 shows the highlighted keyframes that match the ground truth and user summary.

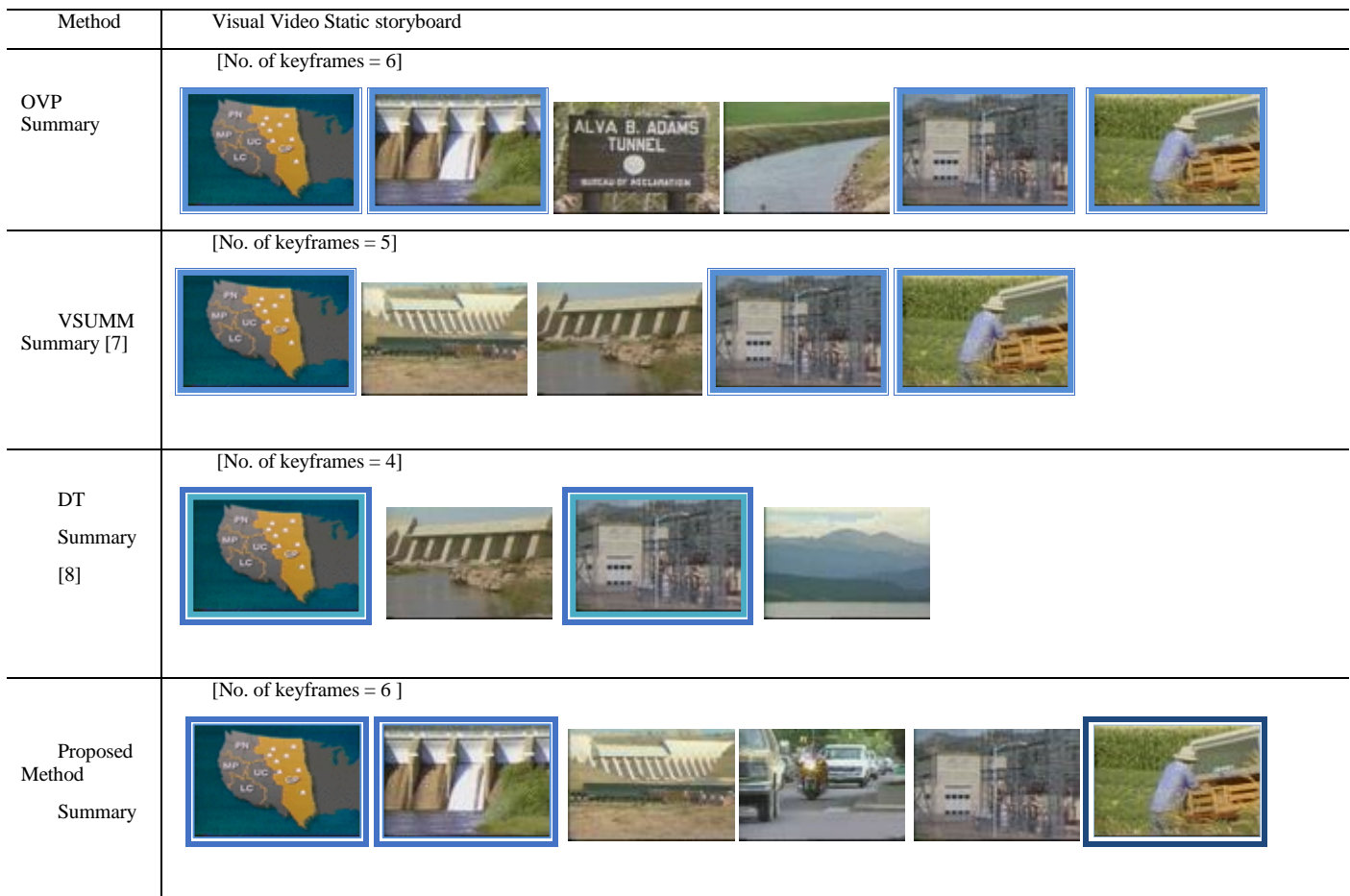


Fig. 7. Comparison of Video Storyboard Created by the Proposed Method Versus DT, VSUMM, and OVP Storyboards.

Table VI shows that the proposed system's performance is better than the DT summary using OVP videos and closer to the VSUMM summary.

TABLE VI. PERFORMANCE COMPARISON OF THE PROPOSED FRACTIONAL ENERGY COEFFICIENTS BASED VIDEO KEYFRAME EXTRACTION METHOD WITH DT[8] AND OVP VIDEO SUMMARIES EXPERIMENTED ON THE OVP DATASET

Name of the Video from (OVP)	Video length in Seconds	Percentage Accuracy (%)		
		VSUMM [7]	DT [8]	Proposed
A New Horizon Segment 5	1.59	78.57	42.86	<b>78.57</b>
A New Horizon Segment 6	1.05	80.00	60.00	60.00
A New Horizon Segment 7	0.47	87.50	75.00	<b>75.00</b>
The Voyage of Lee 15 of 21	1.15	75.00	50.00	<b>83.33</b>
The Voyage of Lee 16 of 21	1.27	71.43	42.86	42.86
The Voyage of Lee 17 of 21	2.25	76.92	61.54	46.15
Average		78.24	55.38	64.32

Comparatively, the VSUMM summaries are closer to the OVP summary, and second next better accuracy is provided by our proposed method of creating a visual storyboard. The proposed storyboard generation method performs better than the Delaunay triangulation method. The same comparison of the proposed method performance is made with VSUMM and DT summaries using SumMe videos.

Table VII shows that the proposed system's performance is better than the DT summary using SumMe input videos.

TABLE VII. PERFORMANCE COMPARISON OF THE PROPOSED FRACTIONAL ENERGY COEFFICIENTS BASED ON VIDEO KEYFRAME EXTRACTION METHOD WITH DT [8] AND SUMME VIDEO SUMMARIES EXPERIMENTED ON SUMME DATASET

Name of the Video (SumMe)	Video length in Seconds	Percentage Accuracy (%)		
		VSUMM [7]	DT [8]	Proposed
Air_Force_One	2.59	68.42	68.42	47.37
Base Jumping	2.38	78.26	47.83	52.17
Bike Polo	1.43	83.33	50.00	55.56
Cooking	1.26	84.62	61.54	<b>76.92</b>
Scuba	1.14	55.56	55.56	<b>77.78</b>
Fire Demo	0.54	66.67	66.67	66.67
Average		72.81	58.33	62.74

So the major contribution of the proposed system is that it overcomes the problem of inclusion of near duplicates due to Delaunay triangulation of clustering. Instead of that proposed system, select the keyframes from all video frames using fractional energy coefficients. In the DT method, summaries are produced for a batch of videos, whereas the proposed system processes each video one by one to include all features giving significance.

## V. CONCLUSION

Video summarization faces the challenge of reducing computational complexity and retrieval accuracy due to its complex nature. This work focuses on reducing visual video frame features for static video summarization, and a new method is proposed for creating a video visual storyboard using transformed visual features with fractional energy coefficients. The transformed coefficients of color planes of the video frames are considered for finding the final feature vector as the set of fractional energy coefficients 25%, 6.25%, and 1.5625% of total coefficients using transforms alias DCT, Slant, Walsh, and Kekre. These features are used for keyframe extraction, and a set of extracted keyframes forms a video visual storyboard.

The average percentage accuracy obtained by the proposed system is 72.51 with the OVP dataset and 73.55 with the SumMe dataset. The keyframes obtained through the proposed system match with the given set of keyframes in the OVP and SumMe dataset videos. The percentage accuracy and matching rate using fractional energy coefficients are higher than using complete 100% energy coefficients used in the existing DT and VSUMM Summary. The keyframe selection done with the proposed use of fractional energy coefficients of transformed video frames for creating a video visual storyboard is better than the use of full energy content, proving the worth of the proposed method. This same method can be further extended for creating the video logs for video storage and indexing as future scope.

## ACKNOWLEDGMENT

I sincerely thank my Ph.D. Guide and mentor Dr. Sudeep D. Thepade (Professor, Computer Engineering, PCCOE Pune and BOS Computer Engineering, Member SPPU Pune) for his valuable guidance in computer vision.

## REFERENCES

- [1] B. T. Truong and S. Venkatesh, "Video abstraction: A systematic review and classification," ACM Trans. Multimed. Comput. Commun. Appl., vol. 3, no. 1, pp. 1–37, 2007.
- [2] K. Schoeffmann and O. Marques, "A Novel Tool for Quick Video Summarization using Keyframe Extraction Techniques A Novel Tool for Quick Video Summarization using Keyframe Extraction Techniques," no. March 2009.
- [3] M. Furini, F. Geraci, M. Montanero, and M. Pellegrini, "STIMO: STill and MOving video storyboard for the web scenario," Multimed. Tools Appl., vol. 46, no. 1, pp. 47–69, 2010.
- [4] M. Gygli, Y. Song, and L. Cao, "Video2GIF: Automatic generation of animated GIFs from video," Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., vol. 2016-Decem, pp. 1001–1009, 2016.
- [5] S. E. F. De Avila, A. Da Luz, A. D. A. Araújo, and M. Cord, "VSUMM: An approach for automatic video summarization and quantitative evaluation," Proc. - 21st Brazilian Symp. Comput. Graph. Image Process. SIBGRAPI 2008, no. June 2014, pp. 103–110, 2008.
- [6] S. E. F. De Avila and A. D. A. Araujo, "VSUMM: An Approach Based on Color Features for Automatic Summarization and a Subjective Evaluation Method," XXII Brazilian Symp. Comput. Graph. Image Process. SIBGRAPI, p. 10, 2009.
- [7] S. E. F. De Avila, A. P. B. Lopes, A. Da Luz, and A. De Albuquerque Araújo, "VSUMM: A mechanism designed to produce static video summaries and a novel evaluation method," Pattern Recognit. Lett., vol. 32, no. 1, pp. 56–68, 2011.

- [8] P. Mundur, Y. Rao, and Y. Yesha, "Keyframe-based video summarization using Delaunay clustering," *Int. J. Digit. Libr.*, vol. 6, no. 2, pp. 219–232, 2006.
- [9] E. J. Y. C. Cahuina and G. C. Chavez, "A new method for static video summarization using local descriptors and video temporal segmentation," *Brazilian Symp. Comput. Graph. Image Process*, pp. 226–233, 2013.
- [10] S. D. Thepade, "Diverse Color Spaces in Video Keyframe Extraction Technique using Thepade's Sorted Ternary Block Truncation Coding with Assorted Similarity Measures," no. GCCT, pp. 256–260, 2015.
- [11] M. Kogler, M. Del Fabro, M. Lux, K. Schöffmann, and L. Böszörményi, "Global vs. Local feature in video summarization: Experimental results," *CEUR Workshop Proc.*, vol. 539, no. December, pp. 108–115, 2009.
- [12] S. D. Thepade and A. A. Tonge, "Extraction of Key Frames from Video using Discrete Cosine Transform," in *International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT)*, 2014, pp. 1294–1297.
- [13] S. SARMADI, "New Approach In Video Summarization Based On Color Feature," vol. 86, pp. 535–542, 2017.
- [14] S. Gupta, "Content-Based Video Retrieval in Transformed Domain using Fractional Coefficients," no. 7, pp. 237–247.
- [15] S. Cvetkovic, M. Jelenkovic, and S. V. Nikolic, "Video summarization using color features and efficient adaptive threshold technique," *Prz. Elektrotechniczny*, vol. 89, no. 2 A, pp. 247–250, 2013.
- [16] X. Yan, S. Z. Gilani, M. Feng, L. Zhang, H. Qin, and A. Mian, "Self-supervised learning to detect key frames in videos," *Sensors (Switzerland)*, vol. 20, no. 23, pp. 1–18, 2020.
- [17] H. Burhan Ul Haq, M. Asif, and M. Bin Ahmad, "Video Summarization Techniques: A Review Article in," *Int. J. Sci. Technol. Res.*, vol. 9, no. 11, pp. 146–153, 2021.
- [18] H. B. Kekre, "Comprehensive Performance Comparison of Cosine, Walsh, Haar, Kekre, Sine, Slant, and Hartley Transforms for CBIR with Fractional Coefficients of Transformed Image," no. 5, pp. 336–351, 2011.
- [19] R. Savran Kızıltepe, J. Q. Gan, and J. J. Escobar, "A novel keyframe extraction method for video classification using deep neural networks," *Neural Comput. Appl.*, vol. 0123456789, 2021.
- [20] V. K. D, S. K. K. L, and J. Majumdar, "Comparison of Video Shot Detection and Video Summarization Techniques," vol. 3, no. 8, pp. 829–833, 2014.
- [21] N. Ejaz, I. Mehmood, and S. Wook Baik, "Efficient visual attention based framework for extracting key frames from videos," *Signal Process. Image Commun.*, vol. 28, no. 1, pp. 34–44, 2013.
- [22] J. Kavitha and P. A. J. Rani, "Static and multiresolution feature extraction for video summarization," *Procedia Comput. Sci.*, vol. 47, no. C, pp. 292–300, 2015.
- [23] A. A. Narvekar, B. E. Student, and B. E. Student, "Color Content-Based Video Retrieval Using Discrete Cosine Transform Applied On Rows and Columns of Video Frames with RGB Color Space," vol. 2, no. 11, pp. 133–135, 2013.
- [24] N. Yadav, "Comprehensive Performance Comparison of Energy Compaction Techniques for CBVR with Transformed Videos," no. 07, pp. 1–7, 2016.

# Denoising of Impulse Noise using Partition-Supported Median, Interpolation and DWT in Dental X-Ray Images

Mohamed Shajahan<sup>1</sup>, Siti Armiza Mohd Aris<sup>2</sup>, Sahnus Usman<sup>3</sup>, Norliza Mohd Noor<sup>4</sup>

Razak Faculty of Technology and Informatics, Universiti Teknologi Malaysia  
Kuala Lumpur Campus, Jalan Sultan Yahya Petra, Malaysia<sup>1,2,3,4</sup>  
University of Business and Technology, Jeddah, Saudi Arabia<sup>1</sup>

**Abstract**—The impulse noise often damages the human dental X-Ray images, leading to improper dental diagnosis. Hence, impulse noise removal in dental images is essential for a better subjective evaluation of human teeth. The existing denoising methods suffer from less restoration performance and less capacity to handle massive noise levels. This method suggests a novel denoising scheme called "Noise removal using Partition supported Median, Interpolation, and Discrete Wavelet Transform (NRPMID)" to address these issues. To effectively reduce the salt and pepper noise up to a range of 98.3 percent noise corruption, this method is applied over the surface of dental X-ray images based on techniques like mean filter, median filter, Bi-linear interpolation, Bi-Cubic interpolation, Lanczos interpolation, and Discrete Wavelet Transform (DWT). In terms of PSNR, IEF, and other metrics, the proposed noise removal algorithm greatly enhances the quality of dental X-ray images.

**Keywords**—Salt and pepper noise; impulse noise; X-ray noise removal; X-ray teeth image quality enhancement; dental X-ray noise reduction

## I. INTRODUCTION

X-ray, CT, and MRI-like medical devices generate many medical images to help doctors. The impulse noise contaminates the medical image pixels, damaging the real data [1]. Currently, X-Ray images are the primary sources for diagnosing dental diseases [2]. The dental diagnosis via dental-X-ray images is a trustful field that provides the exact outcome in dentistry, and it faces the real test of the contingency of impulse noise. The impulse noise is also known as salt and pepper noise. As stated in [3], identifying Dental-images in X-ray is crucial for any intuitive assessment of the denoising of impulse. Impulse noise causes a complete loss of image information, which lowers the quality of X-RAY images [4] [5]. Therefore, impulse noise in dental X-ray images must be addressed. An essential step in any process of enhancing and diagnosing Dental-X-ray images is noise removal [6]. Transmission, acquisition, image processing, and storage are the main causes of impulsive noise. The salt noise has an intensity of 255, whereas the pepper noise has a value of zero [7] [8]. Hence, impulse noise reduction is a crucial part of human dental X-Ray image diagnosis.

A noise removal approach that fixes the corrupted images using either the median or the adjacent pixel value was proposed by Srinivasan et al. [14]. The problem is that they are

unsuitable for low-range noises. Pitas et al. [10] presented a Nonlinear mean filter that works based on the nonlinear-means concept to eliminate impulse noises. The short range of noise reduction is the flaw. Zhang and Li [22] describe an adaptive weighted mean filter (AWMF) to remove impulse noise, especially in images containing high-level noises. The limitation is that the computational time of the work increases in high noise-contaminated images. To sharpen or enlarge the digital image bicubic interpolation is used [31].

Tracey et al. [17] described an enhanced non-local means (NLM) based noise removal technique for X-ray scatter (XBS) images. The edges are not appropriately preserved in this work which is a drawback of this model. Kundu [18] proposes a noise removal method using the PND-NLM technique in X-ray images. The method employs enhancement of X-ray images of bones using Gaussian higher-order derivative operation. X-ray teeth images are not considered in this model, which is a limitation of this work. Shi et al. [16] describe a noise removal technique in real-time X-ray images as it encounters high noise and low contrast. After computing the relationship between noise variance and grayscale value, a modified adaptive noise reduction filter is used to filter the image. Some optimization in work is essential to address the low contrast issue. Shanida et al. [19] suggest a wavelet-based noise removal and enhancement method to address the high noise and low contrast issue in X-Ray teeth images. This work cannot process huge-noise-density teeth X-ray images. Naouel et al. [23] suggest a noise removal and localization method using discrete wavelet transform and thresholding methods. This model extracts seven Regions of Interest from the images, which consumes more time which is a drawback of this work.

Khan et al. [20] developed a Poisson noise and impulse noise removal method based on an improved layer discrimination approach in X-Ray images. The edges are not preserved correctly in this model which is a disadvantage of this work. Markarian et al. [21] propose a denoising method in a compressive sensing framework using high-order total variation method in SAR images. It adopts method based on MAP estimation for denoising and recovering large-size complex SAR images. The thickness of objects in images becomes bulky unnaturally. A noise suppression technique for X-ray images was created by Mandic et al. [24] by altering the intersection of confidence intervals. The relative Intersection of the Confidence Intervals rule is the name of this set of rules.

The teeth X-ray has not been used to examine this work. The existing noise reduction methods suffer from mitigated Peak signal-to-noise ratio, blurred edges, and incapable to resolve a massive range of noises. The proposed NRPMID filter resolves the above issues. It reduces the impulse noise in dental X-ray images. The NRPMID filter is constructed using partition-based Median filter, Interpolation, and DWT techniques. To achieve high denoising accuracy, a new rule set is developed. The partition-based methods are employed according to this new rule set. The biggest drawback that they experience is the noise that is added as a result of the wave's transmitted coherent nature. The image is distorted by these sounds, which might also lead to misidentification. There is a variety of noise in every one of those medical imaging systems. The x-ray images, for instance, are frequently distorted by Poisson noise, salt and pepper noise, and speckle noise. Before continuing with subsequent image processing operations, it is imperative to remove salt-and-pepper noise since this type of noise greatly contaminates images and does severe harm to any knowledge or data that may have been contained in the original image. To overcome certain drawbacks the proposed method is used. The primary contributions of this work are the partition-based methodology and the efficient rule construction.

Section I includes the introduction section. Section II explains the literature review. Section III explains the proposed work, section IV assesses the performance of the NRPMID filter. Section V notifies the conclusion statements about the NRPMID method.

## II. LITERATURE REVIEW

Hawang et al. [9] focused on two algorithms, namely Ranked-order based Adaptive Median Filter (RAMF) and Size based Adaptive Median Filter (SAMF), for denoising the impulse noise. Herein, multiple-size window kernels are used. The surrounding noisy pixels badly impact the denoising quality.

Tao et al. [11] introduced an image enhancement technique to denoise impulse noise based on Tri-state Median Filter (TSMF). Herein, two concepts are used: Standard Median (SM) filter and the Centre Weighted Median (CWM) filter.

Chen and Lien [13] proposed a novel algorithm for removing impulse noise from corrupted images. It employs an efficient impulse noise detector to detect the noisy pixels and an edge-preserving filter to reconstruct the intensity values of noisy pixels.

Gouchol et al. [12] designed a denoising algorithm for impulse noise by performing the Selective removal of impulse noise. This approach works based on homogeneity level information of an image. This work has used the fixed window kernels, which is the disqualification.

Toh et al. [15] described a two-stage noise adaptive technique for eliminating salt-and-pepper noise. The method is based on a fuzzy switching median filter for noise detection and removal. Herein, noisy pixels are determined based on the histogram concept. The artifact generated by the filtering process is found using fuzzy reasoning. The drawback is substantial cost consumption, which has negative effects when used in complicated systems.

## III. METHODOLOGY

The proposed denoising method NRPMID removes impulse noise from Dental X-RAY images using the eight concepts such as Mean filter, Median filter, Bi-linear interpolation-based filter, Partition based Median filter, Bi-cubic interpolation-based filter, Partition based Bi-cubic interpolation-based filter, Partition based Lanczos interpolation-based filter, and Partition based HAAR wavelet transform based filter. This technique makes use of the multi-size window and trimming concepts. This method involves the Trimming concept and multi-size window concept.

Fig. 1 reveals the overall block diagram of the NRPMID filter. The 512x512 size dental grayscale  $I_N$  is fed as input and the final noise-free output is noted as  $I_{NF}$ . The noise-added image is quoted as  $I_N$ . The assignment of  $\alpha = 5$  is made, and  $\alpha$  refers to the maximum permitted iterations. The minimum value of the noisy image is set in  $\beta^1$ . The maximum value of the noisy image is stored in  $\beta^2$ . In this work maximum of five iterations (itr) are used to resolve the noises. If itr=0 then Procedure\_Iteration\_0 is called. If itr=1 then Procedure\_Iteration\_1 is called, and so on. The window size  $\gamma$  is computed as  $\gamma = itr + 1$ . The noisy pixels which meet the condition  $I_N^{ij} \leq \beta^1 \parallel I_N^{ij} \geq \beta^2$ , are stored in  $I_{NI}^{ij} = 0$ , otherwise stored as non-noisy pixels by assigning  $I_{NI}^{ij} = 1$ . Then, the particular iteration is called based on the value of itr-number. Finally the objective function checks for the next progress, by verifying any modification in the current and previous noise-free images.

### A. Itr\_0 Process

The three components Mean computation, Median computation, and Bi-linear interpolation are used in the design of the itr\_0 process. A 3x3 size window is extracted from the noisy locations, and the non-noisy information/intensity is stored in the linear vector  $\delta$ . This scenario is noted as the trimming process. The trimming process yields a maximum of eight elements for a 3x3 window. The non-noise element count is stored in  $TNC$ . If the  $TNC$  is in between one and four, then the mean-oriented denoising is progressed. Otherwise, the standard deviation of  $\delta$  is computed, and if it is less than or equal to 3.2 then the median oriented denoising is progressed, otherwise Bi-linear interpolation [30] oriented denoising is progressed. The denoised outputs are assigned in  $I_{NF}^{ij}$ . Suppose the  $TNC$  is equal to zero then the original intensity is stored in  $I_{NF}^{ij}$ . This process is employed for the entire noisy pixels of the noisy image.

### B. Itr\_1 Process

The Itr\_1 procedure is designed using a new set of rules and four computations viz. Mean, Median, Partition oriented median, and Bi-cubic interpolation. The Itr-0 output is fed as input and quoted as  $I_N$ . A 5x5 window size window is extracted from a noisy pixel, and the trimmed elements are stored in the linear array  $\delta$ . Herein, the maximum numbers of non-noisy elements are 16. The term  $TNC$  is also computed. If  $TNC > 0$  &  $TNC \leq 4$  then Mean oriented denoising is approached. If  $TNC > 4$  &  $TNC \leq 8$  then Median oriented



denoising is approached. If  $TNC > 8$  &  $TNC \leq 12$  then Partition oriented median-based denoising is approached, and it can be calculated by a range of equations from (1) to (4).

$$M_i = FindMedian(\delta) \quad (1)$$

$$M_{P1} = FindMedian(\delta^{0 \text{ to } (\frac{TNC}{2})-1}) \quad (2)$$

$$M_{P2} = FindMedian(\delta^{(\frac{TNC}{2}) \text{ to } TNC-1}) \quad (3)$$

$$M_P = \begin{cases} M_{P1}, & \text{if } abs(M_{P1} - M_i) \leq abs(M_{P2} - M_i) \\ M_{P2}, & \text{otherwise} \end{cases} \quad (4)$$

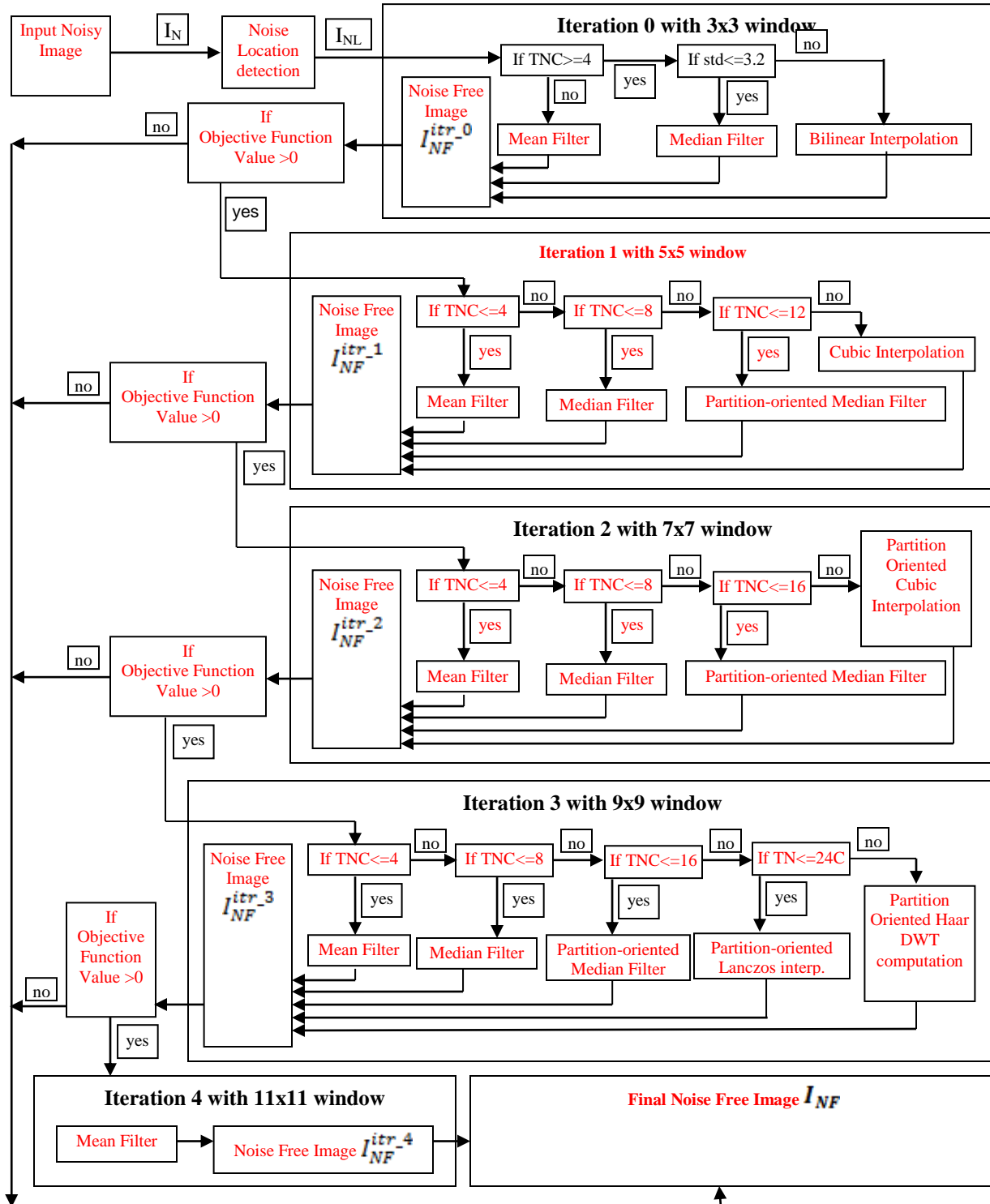


Fig. 1. Working of the Proposed NRPMID Filter.

Where

$M_I$  - Integrated-median,  $M_{P1}$ - Partition\_1 oriented median,  $M_{P2}$ - Partition\_2 oriented median, and  $M_P$  - Final partition-oriented median.

The  $M_I$  is computed from the whole data of  $\delta$ . Herein,  $\delta$  is divided into a couple of divisions, and they are partition\_left and partition\_right. The Partition\_left is used to compute  $M_{P1}$ , and partition\_right is used to calculate  $M_{P2}$ .

The closest part of  $M_I$  is determined via Equation (4) to find the  $M_P$ . If  $TNC > 12$  &  $TNC \leq 16$  then Bi-cubic interpolation-oriented noise removal is approached. The denoised outputs are assigned in  $I_{NF}^{i,j}$ . Suppose the TNC is equal to zero then the original intensity is stored in  $I_{NF}^{i,j}$ . The noisy pixels in the entire noisy image are processed using this method.

### C. Itr\_2 Process

The Itr\_2 procedure is incorporated by a  $7 \times 7$  size neighbor window, new rules, and 4 computations such as Mean, Median, Partition oriented median, and Partition oriented Bi-cubic interpolation. Herein, a maximum of 24 non-noisy neighbors can be extracted. If  $TNC > 0$  &  $TNC \leq 4$ , then mean-oriented noise removal is progressed. If  $TNC > 4$  &  $TNC \leq 8$ , then median value is used for noise cancellation. If  $TNC > 8$  &  $TNC \leq 16$  then Partition oriented median is the source for noise suppression. If  $TNC > 16$  &  $TNC \leq 24$  then Partition oriented Bi-cubic based noise removal is progressed using the range of equations (5) to (8).

$$BC_I = FindBCubicValue(\delta) \quad (5)$$

$$BC_{P1} = FindBCubicValue(\delta^{0 \text{ to } (\frac{TNC}{2})-1}) \quad (6)$$

$$BC_{P2} = FindBCubicValue(\delta^{(\frac{TNC}{2}) \text{ to } TNC-1}) \quad (7)$$

$$BC_P = \begin{cases} BC_{P1}, & \text{if } abs(BC_{P1} - BC_I) \leq abs(BC_{P2} - BC_I) \\ BC_{P2}, & \text{otherwise} \end{cases} \quad (8)$$

where

$BC_I$  - Integrated-BiCubic data

$BC_{P1}$  - Partition\_1 oriented Bi-cubic data.

$BC_{P2}$  - Partition\_2 oriented Bi-cubic data.

$BC_P$  - Final partition oriented Bi-cubic data.

This process is repeated for the whole noisy pixels of the noisy image.

### D. Itr\_3 Process

The Itr\_3 procedure is carried out via a  $9 \times 9$  side window, and the five techniques like i) Mean ii) Median, iii) Partition oriented median, iv) Partition oriented Lanczos interpolation and v) Partition oriented Haar-Wavelet transform. Herein, a maximum of 32 non\_noisy neighbours are collected. The noise reduction is performed based on Fig. 1 corresponding to

iteration\_3. If  $TNC > 16$  &  $TNC \leq 24$  then Partition oriented Lanczos interpolation-based denoising is progressed. Haar\_wavelet transform [32] which is a discrete wavelet transform, is utilized to characterize the image. If  $TNC > 24$  &  $TNC \leq 32$  then Partition-oriented Haar\_wavelet transform-based noise removal is progressed, and it can be done via the range of equations (9) to (12).

$$HW_I = FindHaarWTValue(\delta) \quad (9)$$

$$HW_{P1} = FindHaarWTValue(\delta^{0 \text{ to } (\frac{TNC}{2})-1}) \quad (10)$$

$$HW_{P2} = FindHaarWTValue(\delta^{(\frac{TNC}{2}) \text{ to } TNC-1}) \quad (11)$$

$$HW_P = \begin{cases} HW_{P1}, & \text{if } abs(HW_{P1} - HW_I) \leq abs(HW_{P2} - HW_I) \\ HW_{P2}, & \text{otherwise} \end{cases} \quad (12)$$

where

$HW_I$  - Integrated-HaarWT data.

$HW_{P1}$  - Partition\_1 oriented HaarWT data.

$HW_{P2}$  - Partition\_2 oriented HaarWT data.

$HW_P$  - Final partition-oriented HaarWT data.

This process is continued for the whole noisy pixels of the noisy image.

### E. Itr\_4 Process

The Itr\_4 procedure is designed based on  $11 \times 11$  size window and the Mean computation. The  $\delta$  and TNC are found. If  $TNC > 0$  then mean of  $\delta$  is assigned in the location of noisy data. If  $TNC = 0$  then mean of the whole  $11 \times 11$  size neighbors is assigned as the noise-free data in the corresponding noise location.

In this model, the proposed NRPMID filter reduces the salt and pepper noise or impulse noise in the X-ray dental images.

## IV. DISCUSSION AND ANALYSIS

The MATLAB 15 version was used to build this novel technique. The NMIC Dental X-Ray image dataset [26] and the VAHAB Dental X-Ray image dataset [25] are the databases that were used in this study. 150 dental teeth X-Ray images from each dataset are used in this study as test images. On dental X-ray images, the proposed NRPMID approach is compared to the following three denoising techniques.

- Image denoising using Newton Thiele filter (ID-NTHF)[27].
- Image noise removal based on adaptive fuzzy switching median filter (ID-FSMF) [28].
- Image denoising based on Adaptive Sequentially Weighted Median Filter (ID-AWMF) [29].

The denoising of the proposed technique for 98.32 percent noise corruption is shown in Fig. 2. The input image used in this instance was obtained from the VAHAB database. Fig.

2(a) points out the original image, Fig. 2(b) focuses on the noisy image, and Fig. 2(c) shows the final noise-free image.

1) *Accuracy*: Accuracy is metric used to evaluate how precisely the measured value or findings reflect the real or the original values. The Table I represents the accuracy analysis obtained by the proposed and the other denoising models.

From Fig. 3, it can be observed that the proposed NRPMID filter achieves the highest accuracy value of 0.96% when compared to other models, which indicates that it efficiently reduces the noises from the input image.

2) *Mean Square Error (MSE)*: Mean Square Error is a metric used to measure the amount of error in statistical models. It assesses the average squared difference between the observed and predicted values. When a model has no error, the MSE equals zero. As model error increases, its value increases. The mean squared error is also known as the mean squared deviation (MSD).

In this analysis, the MSE of the proposed NRPMID filter and the other denoising models are analyzed. It is evident from the results demonstrated in Table II and Fig. 4 that the proposed NRPMID filter achieves the lowest MSE value of 0.037.

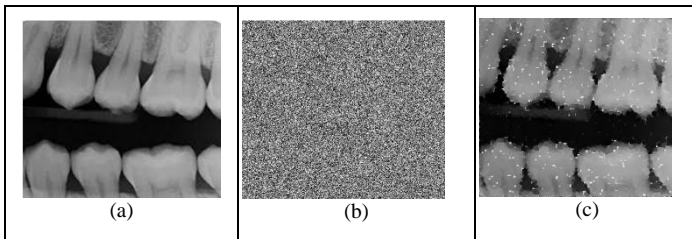


Fig. 2. Noise Removal NRPMID for 98.32% Noise Corruption, (a) Original Input Image (b) Noise Added Image (c) Noise Free Image.

TABLE I. ACCURACY ANALYSIS

Denoising method	Accuracy (%)	
	NMIC database	VAHAB database
ID-NTHF	0.74	0.76
ID-FSMF	0.77	0.79
ID-AWMF	0.89	0.90
NRPMID	0.95	0.96

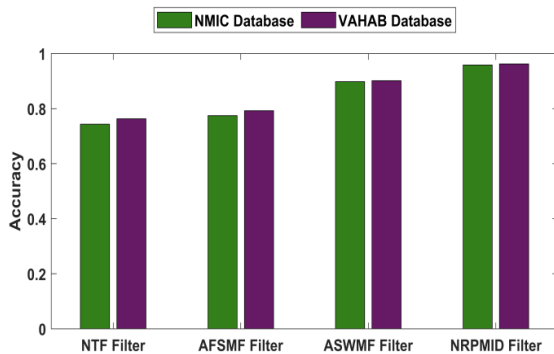


Fig. 3. Accuracy Analysis on Four Methods and Two Datasets.

TABLE II. MSE ANALYSIS

Denoising method	MSE	
	NMIC database	VAHAB database
ID-NTHF	0.256	0.236
ID-FSMF	0.225	0.207
ID-AWMF	0.101	0.098
NRPMID	0.041	0.037

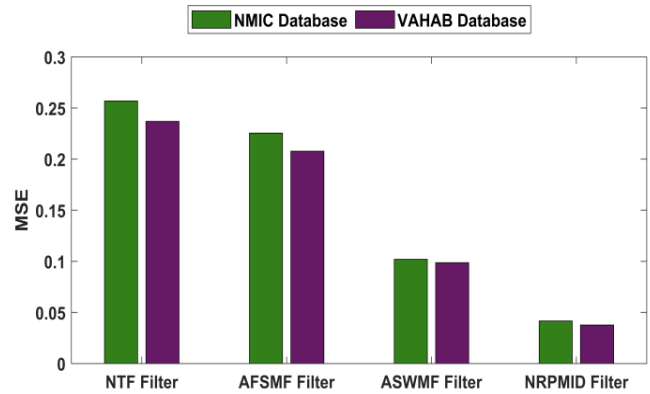


Fig. 4. MSE Analysis for Four Methods and Two Datasets.

3) *Average MSE analysis*: The average MSE analysis is also computed in this analysis and results are represented in Table III.

TABLE III. AVG. MSE ANALYSIS FOR 90% NOISE CORRUPTION

Denoising method	Avg. MSE	
	NMIC database	VAHAB database
ID-NTHF	656.276	629.623
ID-FSMF	591.679	533.435
ID-AWMF	427.640	391.810
NRPMID	131.555	124.198

The proposed NRPMID filter's effectiveness in the average MSE analysis case with noise corruption of 90% is shown in Table III. When compared to other approaches, the NRPMID method produces low MSE values.

4) *PSNR Analysis*: The Peak Signal to Noise Ratio (PSNR) analysis is computed and the High PSNR means better denoising.

The average PSNR values with noise corruption of 90% can be seen in Fig. 5. Each database's 150 images are used to compute it. It demonstrates that the NRPMID filter can be used to restore more successfully than the conventional filters.

5) *Root Mean Square Error (RMSE)*: The mathematical derivation for computing RMSE is provided in Equations (13).

$$Error_{RMSE} = \sqrt{\frac{\sum_{i=1}^N (p_i - r_i)^2}{T}} \quad (13)$$

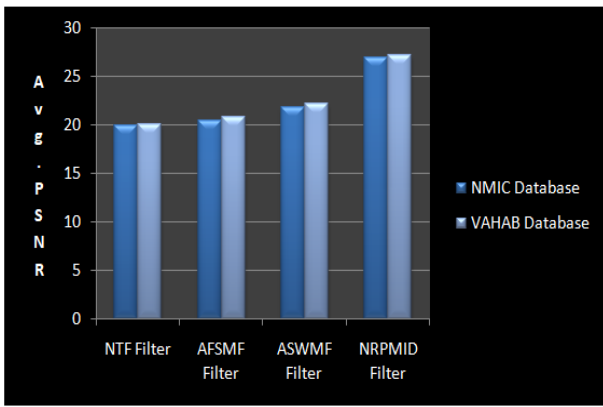


Fig. 5. Average PSNR Analysis for Noise Corruption of 90%.

TABLE IV. RMSE ANALYSIS

Denoising method	MSE	
	NMIC database	VAHAB database
ID-NTHF	0.506	0.486
ID-FSMF	0.474	0.455
ID-AWMF	0.319	0.314
NRPMID	0.203	0.194

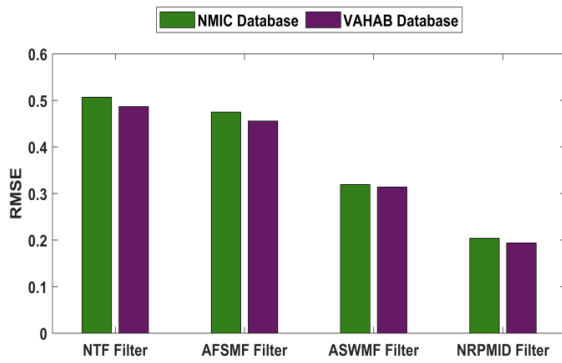


Fig. 6. RMSE Analysis.

The RMSE analysis of the proposed NRPMID filter and other denoising models are expressed in Table IV and Fig 6. From the results, it is evident that the proposed work achieves the lowest RMSE value of 0.194 which indicates the accuracy of the NRPMID filter is high.

6) *Structural Similarity Index (SSIM)*: A perceptual metric called the Structural Similarity Index (SSIM) measures the loss of image quality put on by denoising [33]. The denoising Quality is better if the MSSIM value is high.

The average SSIM results in 60% noise corruption, as shown in Fig. 7. It illustrates that the proposed NRPMID approach obtains higher SSIM values than the traditional methods.

7) *Average IEF analysis*: The IEF analysis is used to assess the effectiveness of salt and pepper noise denoising in dental X-ray images. Better denoising quality is correlated with higher IEF and vice versa [34].

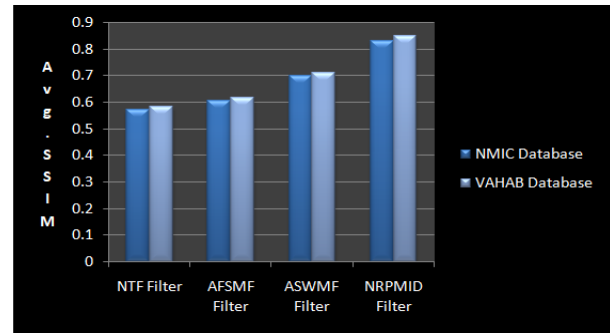


Fig. 7. Avg. SSIM Analysis for 60% Noise Corruption.

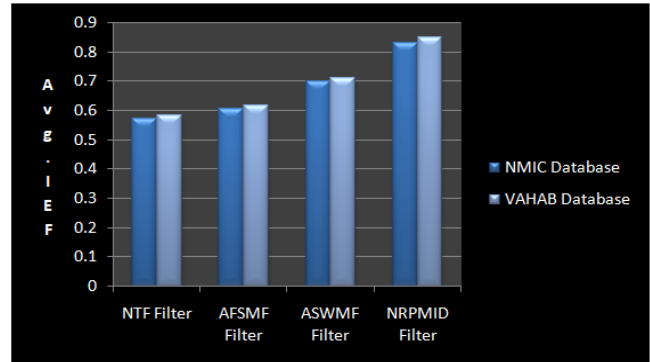


Fig. 8. Avg. IEF Analysis for 90% Noise Corruption.

The average IEF analysis values for noise corruption of 90% are shown in Fig. 8. The proposed NRPMID filter outperforms existing techniques in terms of high IEF values. The VAHAB database is the finest database to support the proposed work, according to the analysis.

## V. CONCLUSION

This research proposes the NRPMID filter to reduce the impulse noise in teeth X-ray medical images. The NRPMID filter can resolve salt and pepper noise from the dental X-Ray images, which are even corrupted by 98.32%. The proposed filter improves the visual quality of dental X-ray images. It assists dental pathologists in having an accurate diagnosis. The time complexity of this approach is significantly less. The analysis section proves the noticeable results of the NRPMID filter than the existing denoising methods. The advantage of the method is simple and less costly. The drawback of the proposed system is the high computational time. In the future, this research can be extended to solve the Poisson noise in dental X-Ray images. Also, image enhancement techniques will be used to enhance the image quality.

## REFERENCES

- [1] Sumitra, P. "A comparative study algorithm for noisy image restoration in the field of medical imaging." International Journal of Advanced Information Technology (IJAIT) 6.1 (2016): 35-42.
- [2] Kadam, Chaitali, and P. S. B. Borse. "A Comparative Study of Image Denoising Techniques for Medical Images." image 4.06 (2017).
- [3] K. Shanida, R. Shayini, and C.S. Sindhu, "Dental image enhancement using wavelet decomposition and reconstruction", International Journal of Recent Advances in Engineering & Technology (IJRAET), vol. 4, issue 7, 2016.
- [4] HosseinKhani, Zohreh, et al. "Adaptive real-time removal of impulse noise in medical images." Journal of medical systems 42.11 (2018): 1-9.

- [5] Nitu Kumari, Kusum Kumari, Manisha Tigga and Sushanta Mahanty, "Noise detection and noise removal techniques in medical images", International journal of computer engineering and applications, vol. 10, special issue, 2016.
- [6] K. Shanida, R. Shayini and C.S. Sindhu, "Dental image enhancement using wavelet decomposition and reconstruction", International Journal of Recent Advances in Engineering & Technology (IJRAET), vol. 4, issue 7, 2016.
- [7] HosseinKhani, Zohreh, et al. "Adaptive real-time removal of impulse noise in medical images." Journal of medical systems 42.11 (2018): 1-9.
- [8] Nitu Kumari, Kusum Kumari, Manisha Tigga and Sushanta Mahanty, "Noise detection and noise removal techniques in medical images", International journal of computer engineering and applications, vol. 10, special issue, 2016.
- [9] Khan, Sajid, and Dong-Ho Lee. "An adaptive dynamically weighted median filter for impulse noise removal." EURASIP Journal on Advances in Signal Processing 2017, no. 1 (2017): 1-14.
- [10] Kandemir, Cengiz, Cem Kalyoncu, and Önsen Toygar. "A weighted mean filter with spatial-bias elimination for impulse noise removal." Digital Signal Processing 46 (2015): 164-174.
- [11] Thanh, Dang Ngoc Hoang, and Serdar Enginoğlu. "An iterative mean filter for image denoising." IEEE Access 7 (2019): 167847-167859.
- [12] P. Gouchol, L. Jyh-charn and N.A. Sanju, "Selective removal of impulse noise based on homogeneity level information", IEEE Transactions on image processing, vol. 12, no. 1, pp. 85-92, 2003.
- [13] Chen PY, Lien CY. An efficient edge-preserving algorithm for removal of salt-and-pepper noise. IEEE Signal Process Lett. 2008;15(2):833-6.
- [14] K.S. Srinivasan and D. Ebenezer, "A new fast and efficient decision-based algorithm for removal of High-Density impulse noises", IEEE Signal processing letters, vol. 14, no. 3, pp. 189-192, 2007.
- [15] K.K.V. Toh and N.A.M. Isa, "Noise Adaptive Fuzzy Switching Median Filter for Salt and Pepper noise reduction", IEEE Signal Processing letters, vol. 17, no. 3, pp. 281-284, 2010.
- [16] H. Shi, J. Shao, D. Du, B. Chang and H. Cao, "Noise Reduction of the Real-time X-ray Image Based on Modified Adaptive Local Noise Reduction Filter", IEEE, 4th International Congress on Image and Signal Processing, 2011.
- [17] B.H. Tracey, E.L. Miller, M. Schiefele, C. Alvino and O.A. Kofahi, "Denoising approaches for X-ray personnel screening systems", IEEE Conference on Technologies for Homeland Security (HST), 2012.
- [18] Kundu R, "Structural Enhancement of Digital X-ray Image of Bone with a Suitable Denoising Technique", Indian conference on medical informatics and telemedicine (ICMIT), 2013.
- [19] K.Shanida, R. Shayini and C.S. Sindhu, "Dental Image Enhancement Using Wavelet Decomposition and Reconstruction". International Journal of Recent Advances in Engineering & Technology (IJRAET), vol. 4, issue 7, pp. 2347 -2812, 2016.
- [20] S.U. Khan, M. Ishaq, N. Ullah, A. Ahamd and I. Ahmed, "A novel algorithm for removal of noise from X-Ray images", International Journal of Computer Science and Information Security, vol. 14, issue 10, 103-109, 2016.
- [21] H. Markarian, and S. Ghofrani, "High-TV based CS framework using MAP estimator for SAR image enhancement", IEEE journal of selected topics in applied earth observations and remote sensing, vol. 10, issue 9, pp. 4059-4073, 2017.
- [22] Zhang P, Li F. A new adaptive weighted mean filter for removing salt-and-pepper noise. IEEE Signal Process Lett. 2014; 21(10):1280-3.
- [23] G. Naouel, M.C. Olfa, M. Mokhtar and M. Jerome, "Evaluation of DWT denoise method on X-ray images acquired using flat detector", IEEE 4th Middle East Conference on Biomedical Engineering (MECBME), 2018.
- [24] I. Mandic, H. Peic, J. Lerga and I. Stajduhar, "Denoising of X-ray images using the Adaptive Algorithm Based on the LPA-RICI Algorithm", Journal of imaging, vol. 4, issue 34, pp. 1-15, 2018.
- [25] Vahab database: Accessed from <https://mynotebook.labarchives.com/share/Vahab/MjAuOHw4NTc2Mi8xNi9UcmVITm9kZS83Nm50Tk2MDZ8NTUOA>, Accessed on [25-Mar-2021].
- [26] NMIC database: Accessed from: <https://data.mendeley.com/datasets/hxt48yk462/1>, Accessed on [25-Mar-2021].
- [27] Tian Bai and Jieqing Tan, "Automatic detection and removal of high-density impulse noises", IET image processing, vol. 9, Issue 2, pp. 162 - 172, 2015.
- [28] Ayyaz Hussain and Muhammad Habib, "A new cluster-based adaptive fuzzy switching median filter for impulse noise removal", SPRINGER, Multimedia Tools and applications, vol. 76, pp. 22001-22018, 2017.
- [29] Jiayi Chen, Yinwei Zhanz and Huying Cao, "Adaptive Sequentially Weighted Median Filter for Image Highly Corrupted by Impulse Noise", IEEE Access, vol. 7, pp. 158545 - 158556, 2019.
- [30] Bi-linear interpolation, Accessed from <https://theailearner.com/2018/12/29/image-processing-bilinear-interpolation/>, Accessed on [4-Apr-2021].
- [31] Bi-cubic interpolation, Accessed from <https://medium.com/hd-pro/bicubic-interpolation-techniques-for-digital-imaging-7c6d86dc35dc>, Accessed on [4-Apr-2021].
- [32] Haar Transform, Accessed from <https://en.wikipedia.org/wiki/ Haar\_wavelet>, Accessed on [6-Apr-2021].
- [33] SSIM, Accessed from <https://www.imatest.com/docs/ssim/>, Accessed on [5-Apr-2021].
- [34] V. Jayaraj and D. Ebenezer, "A New Switching-Based Median Filtering Scheme and Algorithm for Removal of High-Density Salt and Pepper Noise in Images", EURASIP Journal on Advances in Signal Processing, Vol. 2010, pp. 1-11. 2010.

# An End-to-End Big Data Deduplication Framework based on Online Continuous Learning

Widad Elouataoui<sup>1</sup>

Laboratory of Engineering Sciences  
National School of Applied Sciences, Ibn Tofail University  
Kenitra, Morocco

Saida El Mendili<sup>3</sup>

Laboratory of Engineering Sciences  
National School of Applied Sciences, Ibn Tofail University  
Kenitra, Morocco

Imane El Alaoui<sup>2</sup>

Telecommunications Systems and Decision Engineering  
Laboratory, Ibn Tofail University  
Kenitra, Morocco

Youssef Gahi<sup>4</sup>

Laboratory of Engineering Sciences  
National School of Applied Sciences, Ibn Tofail University  
Kenitra, Morocco

**Abstract**—While big data benefits are numerous, most of the collected data is of poor quality and, therefore, cannot be effectively used as it is. One pre-processing the leading big data quality challenges is data duplication. Indeed, the gathered big data are usually messy and may contain duplicated records. The process of detecting and eliminating duplicated records is known as Deduplication, or Entity Resolution or also Record Linkage. Data deduplication has been widely discussed in the literature, and multiple deduplication approaches were suggested. However, few efforts have been made to address deduplication issues in Big Data Context. Also, the existing big data deduplication approaches are not handling the case of the decreasing performance of the deduplication model during the serving. In addition, most current methods are limited to duplicate detection, which is part of the deduplication process. Therefore, we aim through this paper to propose an End-to-End Big Data Deduplication Framework based on a semi-supervised learning approach that outperforms the existing big data deduplication approaches with an F-score of 98,21%, a Precision of 98,24% and a Recall of 96,48%. Moreover, the suggested framework encompasses all data deduplication phases, including data pre-processing and preparation, automated data labeling, duplicate detection, data cleaning, and an auditing and monitoring phase. This last phase is based on an online continual learning strategy for big data deduplication that allows addressing the decreasing performance of the deduplication model during the serving. The obtained results have shown that the suggested continual learning strategy has increased the model accuracy by 1,16%. Furthermore, we apply the proposed framework to three different datasets and compare its performance against the existing deduplication models. Finally, the results are discussed, conclusions are made, and future work directions are highlighted.

**Keywords**—Big data deduplication; online continual learning; big data; entity resolution; record linkage; duplicates detection

## I. INTRODUCTION

Nowadays, data quality is gaining wide attention from both academics and professionals. Indeed, data quality dramatically impacts the business as executives rely mainly on data to manage their business and make informed decisions

[1]. Indeed, better data quality translates directly into better business value. Data quality could be defined in terms of different dimensions such as completeness, accuracy, timeliness and consistency [2] [3].

This article addresses one of data quality's main aspects: uniqueness. Uniqueness ensures that there is only one instance of each information in the dataset and thus points out that there should be no data duplicates [4]. Indeed, data duplication issues are not only related to storage. Duplicate data also lead to inaccurate analysis, which may cause significant problems and costly mistakes. There are many sources of data redundancy, including users providing erroneous information, typing errors, data integration, etc. With the emergence of Big Data, data duplication has become more common and challenging. This is related to big data Volume, Variety, Velocity, and other characteristics of big data known as Big Data V's [5] [6]. Thus, because of the particular characteristics of big data, new data deduplication issues were raised related to the huge data volume, variety of data sources, inconsistency of data types and schemas, and so on.

Therefore, duplicate detection approaches have been widely discussed in the literature under different names, such as entity resolution, deduplication, or record linkage. All these terms refer to the same meaning: identifying records referring to the same real-world entity. The deduplication process is usually followed by an entity consolidation or fusion process defining the unified representation of duplicated values that best represents the real-world entity. Even if data deduplication was widely discussed in the literature, more efforts are needed to address the challenges related to Big Data Deduplication. Indeed, most existing big data deduplication approaches focus only on data volume. Also, most existing methods are limited to the duplicate detection phase, which is only a part of the deduplication process. Moreover, the current deduplication approaches are not ensuring a maintained accuracy score during the serving, so the model's performance usually decreases over time [7].



Believing that Big Data Deduplication should be addressed more comprehensively, we aim through this paper to enhance big data quality measurement with three main contributions:

- We suggest an End-to-End Big Data Deduplication Framework encompassing five phases: data pre-processing, data labeling, duplicate detection, data cleaning, and finally, model monitoring using continual retraining.
- We address the issue of the decreasing performance of the deduplication model by setting an online learning strategy for big data deduplication to maintain a high accuracy level during the serving.
- We design a framework that outperforms the existing big data deduplication methods and provides the best results based on a Semi-Supervised learning approach.

The rest of this paper is organized as follows: Section 2 describes the research methodology followed for the literature review. Section 3 reviews the most recent and relevant studies that have tackled data deduplication. Section 4 highlights the importance of deduplication for big data. Section 5 presents our suggested end-to-end big data deduplication framework. Section 6 offers the implementation of the suggested framework and discusses the obtained results. Finally, we highlight the primary outcomes as well as some research outlooks.

## II. RESEARCH METHODOLOGY

A systematic literature review was conducted to capture and synthesize the relevant and available studies addressing data quality measurement. This literature review was performed following the guidelines stated in [8], where the authors have proposed a review methodology that consists of planning the review by preparing a review proposal. A second step consists of searching and selecting studies. Finally, the main findings of the review are reported. The goal of this study was to choose two main kinds of contributions:

- Studies suggesting deduplication frameworks in a big and non-big data context.
- Studies addressing Uniqueness as a quality metric

For this, primary research was conducted first using generic keywords such as “Data Deduplication”, “Entity Resolution” and “Data Uniqueness”. Then, to capture studies about big data, specific keywords such as “Big data Deduplication”, “Big Data Entity Resolution”, and “Big Data Redundancy” were used. Then, abstracts were reviewed, and irrelevant papers were excluded. This primary search yielded 60 articles. The research was limited to recent articles published in journals and conference proceedings and was performed on: IEEE Xplore, Springer, Google Scholar, Science Direct, Research Gate, and ACM digital libraries. After a literature search, the next step consists of narrowing down the papers based on their relevancy, freshness, and availability. For this, a diagonal reading was performed on the selected papers filtered out based on multiple criteria: we included studies that were addressing data deduplication, recent, available, in English, and published in digital libraries.

A total of 23 articles were selected, followed by a more in-depth analysis.

Further, we reviewed the references of the selected studies and added two more articles to the selected papers. Then, the chosen studies were thoroughly read and carefully examined, and 17 studies were deemed relevant to the scope of our research. Finally, the articles' descriptive details were checked and filed in a Zotero database. This literature review has shown a significant lack of deduplication frameworks that fit big data requirements, which motivates us to perform an in-depth analysis of the current state of the art to frame the need and make a significant contribution. The following section reviews the papers selected for our study and highlights the main findings of this literature review.

## III. RELATED WORK

Data deduplication is a trending topic that many researchers have long addressed in the literature. Thus, many approaches for data deduplication have been suggested in the literature, such as [9], where the authors have proposed a six-step deduplication framework that detects duplicated records using record linkage. The framework includes preparing data, matching attributes using sorted neighborhood, building a decision mode, and clustering. In this paper, the authors have raised the current issue of the lack of labeled data for big data deduplication, which hinders the evaluation of the model performance. This issue was also mentioned in [10], where the authors have raised the lack of labeled data for deduplication and suggest using active learning as an alternative. The authors have achieved the highest results with an F-score of 98,4% for structured datasets. However, for dirty datasets, a lower score of 52% was achieved. Deep learning was also used to address data duplication, such as in [11], where the authors define a binary classification approach for safety engineering based on fuzzy string-matching algorithms. The proposed approach is based on Convolutional Neural Networks (CNN) binary classifier and string similarity-based classifier. All the research mentioned above has significantly contributed to duplicate detection and entity resolution. However, these approaches are not appropriate for large-scale datasets in terms of accuracy and execution time. Also, in addition to the dataset volume concern, the above approaches did not address the particular issues raised in a big data context. Indeed, with the emergence of big data, new challenges have been raised, such as the diversity of data sources, the variety of data types, and the high velocity and veracity of data [12] [13]. These particular issues were discussed in recent studies, such as in [14], where the authors have performed a survey of the indexing techniques for big data deduplication. The experiments have shown that sorted neighborhood is the best indexing technique for large datasets in terms of complexity. Also, in [7], Christophides et al. have performed a comprehensive survey of all the existing methods for entity resolution. They provided an overview of the different steps of entity resolution for big data, including blocking, block processing, matching, and clustering. The authors have also raised the challenge of the decreasing performance of the deduplication methods over time. Likewise, in [15], the authors have discussed big data's challenges to entity resolution and proposed a hybrid

similarity measurement approach based on traditional syntactic and word-embedding approaches. In [16], Abd El-Ghafar et al. have suggested an entity resolution approach for big data based on hashing TF and Jaccard similarity. The approach was applied to seven scenarios where different Natural Language Processing (NLP) techniques were used to show the impact of these techniques on entity resolution. This approach reaches an accuracy of 91% for a dataset of 1M records. To address data duplication in the web of data, Efthymiou et al. have proposed in [17] a deduplication that allows reducing the required number of pairwise comparisons. The suggested process blocks data when comparing entity descriptions within the same blocks. The results show a high performance of the suggested method; however, it is only appropriate for the web of data as it is based on entity descriptions. Using deep learning, the authors in [18] have introduced a new Stacked Dedupe Learning entity resolution approach based on Bidirectional Recurrent Neural Networks (RNNs) with Long Short-Term Memory (LSTM). However, the study does not show the impact of big data on the performance and accuracy of the model. Moreover, recent studies have addressed the data blocking for big data, such as in [19], where the authors have defined a progressive blocking (PB), detecting 93% of duplicates during the first third of the execution time. Likewise, a multi-phase blocking strategy detecting big data duplicates has been suggested in [20].

Even if data deduplication has been widely discussed in the literature, there have been few efforts to address deduplication issues in the big data context. In addition, most of the existing approaches are only limited to the duplicate detection phase and are not comprehensively managing data deduplication. Furthermore, the model's predictive performance usually degrades over time as the data keep changing. To the best of our knowledge, no deduplication approaches have been set to address the decreasing accuracy during the model serving. Therefore, we aim through this paper to propose an end-to-end Big Data Deduplication framework with three main contributions:

- Ensuring increased performance and accuracy of the deduplication by setting an online learning strategy for deduplication.
- Setting a more comprehensive Big Data deduplication Framework that addresses all the big data deduplication processes and consists of five steps: data preprocessing, data labeling, duplicate detection, data cleaning, and model monitoring using continual retraining.
- Suggesting a novel framework that outperforms the existing big data deduplication methods based on a Semi-Supervised learning approach.

In the next section, we highlight the importance of deduplication for big data.

#### IV. BIG DATA DEDUPLICATION

In a Big Data context, ensuring data quality has always been a critical concern for data managers. Data quality could be defined in terms of multiple dimensions, also called “Data

Quality Dimensions” such as completeness, accuracy, readability, consistency, etc. In this paper, we are addressing one of the primary data quality dimensions: uniqueness. It refers to the unicity of the information provided by the dataset and ensures that there are no duplicated records. To improve data uniqueness, data should be cleaned from duplicated data. This process is known as Data Deduplication. Data duplication can occur for different reasons, such as data integration, where data are gathered from multiple data sources so the same information can be recorded more than once in another format [21] [22]. Also, data duplication could be related to human errors, so the same person, for example, could provide data with slightly different information intentionally or by mistake multiple times. Indeed, Experian [23] found that human input error is the leading cause of data inaccuracy and duplication. Data duplication heavily impacts data analysis and can negatively affect the business. Data duplication can bias data analytics. For example, companies lack a single customer view with duplicated customer dataset. They could not have a clear idea about the real number of their customers and their behavior which may hinder activities like targeted marketing. Also, data duplication incurs a high cost as it leads to wasteful marketing activities, such as targeting the same customer multiple times. Data duplication could also be costly in terms of storage, as redundant records can take up a lot of space, which increases storage costs. A recent study [24], about the impact of data duplication has shown that companies that store big data and apply a backup policy can see that 80% of their corporate data are duplicated. Also, according to another study [25], reducing the transmitted data can save money in terms of storage costs and backup speed up to 50%. Thus, data deduplication helps optimize marketing spending in terms of time and cost. In short, data duplication can result in significant damage and cost for businesses and, therefore, should be addressed effectively for accurate and successful data management. In the next section, we present the suggested end-to-end big data deduplication framework and describe each step of the framework straightforwardly.

#### V. A SMART END-TO-END BIG DATA DEDUPLICATION FRAMEWORK

In this section, we present an end-to-end Big Data deduplication Framework, shown in Fig. 1 to 6 that consists of five steps: The first step is a preprocessing phase where data is cleaned and prepared for deduplication due to the low quality of the extracted data in big data environments. The next step consists of building a training dataset using an automated data labeling process. Then, fuzzy matching is performed on the dataset to detect duplicates. The detected duplicates are then cleaned using the appropriate strategies. Finally, the model is deployed using a real-time continual learning strategy for continuous accuracy improvement during the serving. The framework is designed to address the different issues linked to big data environments. In the following, we provide a detailed description of each stage of the framework.

##### A. Pre-processing

Because of the Big Data V's, the extracted data in big data environments are usually unstructured, noisy, and poorly formatted. Therefore, going through a pre-processing phase is

highly required before using data [26]. In this first phase, raw data is prepared and converted into a more appropriate format making it understandable and suitable for use by Machine Learning (ML) algorithms. This process significantly impacts the efficiency and accuracy of the model and can ruin the subsequent phases if it is not done correctly. In the following, we present the transformations required to prepare big data for deduplication, as shown in Fig. 2.

1) *Feature Selection and Extraction*: Feature Selection and Extraction are crucial in dealing with a high-dimensional dataset as not all the extracted data in Big Data environments are relevant for the intended use. The goal is to keep relevant information by selecting only the most informative variables (Feature Selection) or creating new useful ones (Feature Extraction). This process is required for data deduplication as it allows determining the most significant features on which the model will be based to detect duplicates.

2) *Imputing*: Big data is usually messy, skewing data analysis and leading to biased results. Imputing data is required for deduplication, especially when there is a large number of missing values, as, with low information, duplicates cannot be detected effectively. There are various ways to address the missing values depending on the ratio of the missing values. The missing values can either be ignored, deleted, or replaced by an estimate based on the existing part of the data. The estimated value could be the mean value, the most frequent value, the min or max value, etc. Data could also be attributed using ML algorithms such as K-Nearest Neighbour and Multivariate Imputation or deep learning such as DataWig.

3) *Encoding*: Encoding is the process of converting categorical variables into numeric types. Most ML algorithms cannot handle absolute values and work better with numerical

inputs. There are multiple techniques for encoding, such as Label Encoding, One Hot Encoder, Vector Indexer, etc. Moreover, encoding ensures data consistency, a crucial factor for data deduplication. Indeed, as big data are gathered from multiple sources, categorical values may be represented differently, such inconsistency issues impede duplicate detection.

4) *Upper casing/Lower casing*: This transformation consists of standardizing text data to all Lowercase or Uppercase. For the sake of simplicity, it is more common to convert all data to lowercase, especially for NLP applications. This process is also essential for deduplication as the same word (Good/ good) may be taken as different words (in the vector space model) if we ignore this transformation.

5) *Stop Words and Symbols Removal*: This process consists of removing irrelevant words- the most common words- from the text data. The idea behind eliminating stop words is to provide more importance to the information contained within data, as ignoring them doesn't drastically impact the meaning. Also, the dataset should be cleaned from special symbols and punctuations as they will not help identify similarities. According to the context, other text elements could be removed, such as URLs, HTML tags, etc.

6) *Normalization*: Due to the variety of data sources, some variables in the data may have different scales. This inconsistency at the scale level will bias duplicate detection as records should be compared based on a unified scale. To overcome this, data should be normalized so that the range of all the variables is similar (usually between 0 and 1).

We consider these transformations the most important ones to prepare data for deduplication. However, according to the dataset context, more text cleaning may be needed, such as Spell Corrections and Stemming.



Fig. 1. End-to-End Data Deduplication Approach.

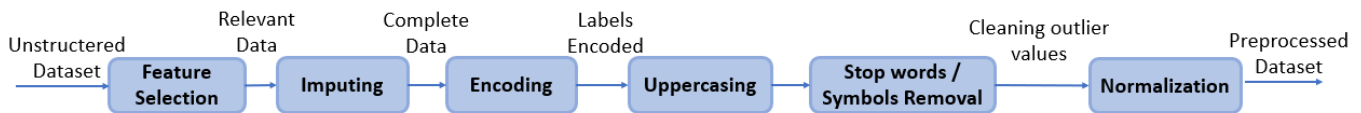


Fig. 2. Preprocessing Steps.



Fig. 3. Labeling Steps.

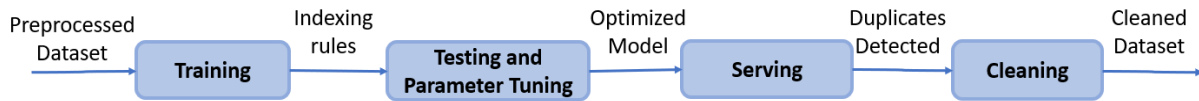


Fig. 4. Deduplication and Cleaning Steps.

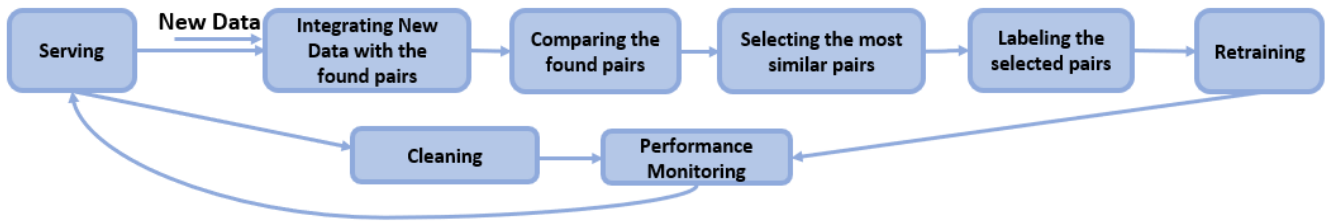


Fig. 5. Online Continual Learning Process.

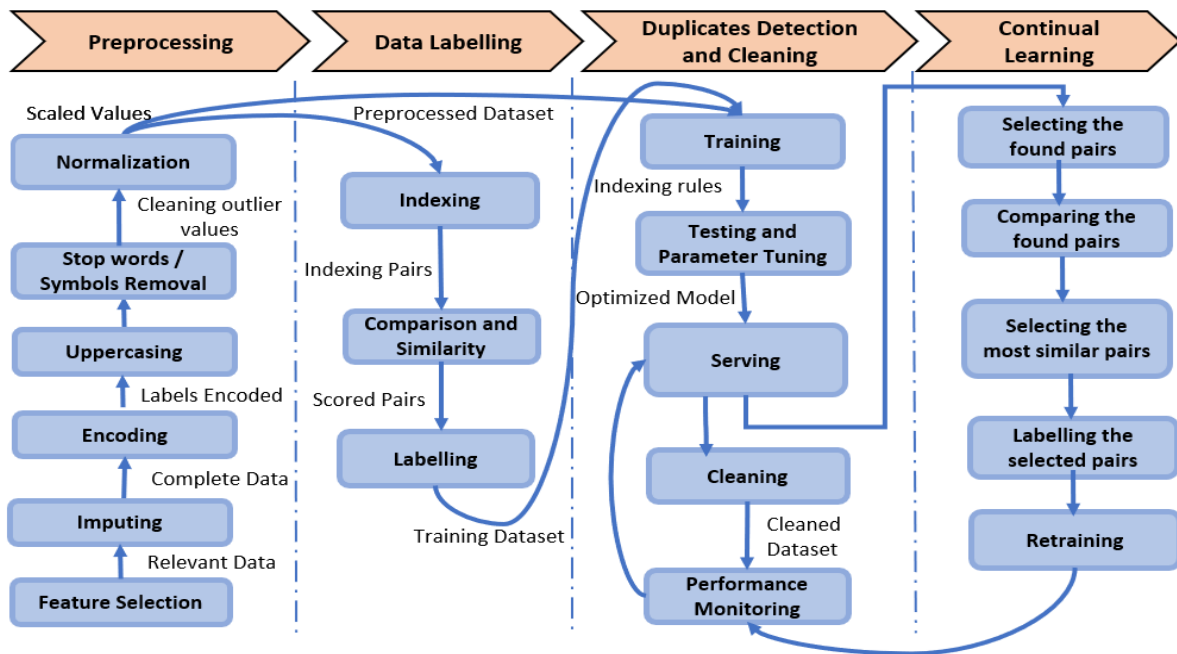


Fig. 6. Data Deduplication Pipeline.

### B. Data Labeling

Once the dataset is pre-processed, and in a ready-to-use state, the next step consists of building a labeled dataset that will be used to train the deduplication model. Labeling data is one of the most challenging tasks that could be faced in AI projects. According to a study [27], labeling data takes up to 80% of AI project time. If most labeling approaches use human labelers, this solution becomes unsuitable when dealing with big data, not only for quantity reasons but also for quality reasons. To overcome this, we are using an automated approach to produce labeled data for deduplication based on Record Linkage techniques. This approach, shown in Fig. 3, consists first of indexing records into pairs. Then, a weighted similarity score is computed to determine if the couples are duplicates, and finally, pairs are labeled based on their similarity score.

1) *Indexing*: Indexing consists of generating pairs of candidate records. The idea behind this step is not to create all

possible combinations of record pairs in the data set, as it will lead to quadratic time complexity, but to select only the likely duplicated pairs. Several indexing techniques are available for record linkage, such as Blocking, Sorted Neighbourhood, TF-IDF, etc. In this paper, we use Sorted Neighbourhood for pairs indexing as it is the most suitable indexing technique for big data [14]. More details are provided in the implementation section.

2) *Comparison and Similarity*: After generating the record pairs, a comparison of the candidate records is performed, and a similarity score is then attributed to each pair. Depending on the field type (string, numerical value, date...), multiple comparison measures, such as Jarowinkler, Levenshtein, Cosine, Jaccard, etc., could be used. For more accurate measurements, weights could also be assigned to data fields, as some areas may be more significant than others to determine duplicate records.

3) *Labeling*: Once similarity scores are measured, pairs are classified using supervised or unsupervised methods such as Optimal Threshold, SVM, K-Means, Farthest First, etc. Pairs are then classified into two classes (matches/non-matches). As a record can have more than one duplicate, we are suggesting in this approach to gather duplicates into clusters instead of pairs so that each cluster can contain more than two records. Non duplicate records are removed, and only matching records are kept as a training dataset.

### C. Duplicates Detection

With a set of labeled data, we can thus train the deduplication model. Then, use the trained model to identify matches and find the correct parameters to get optimal results. It is worth noting that in this approach duplicate detection is not based on a text comparison but on deduplication predicates and indexing rules generated by the model after the training. More details about the generated indexing rules are provided in the implementation section.

1) *Training*: This first step consists of training the model to classify records as duplicates and non-duplicates based on the training dataset. At the end of the training, the model comes up with indexing rules that will be used to identify potential matches. Thus, records will be blocked by matching the deduced indexing rules (also called Predicates) during the learning.

2) *Testing and Parameter Tuning*: The next step consists of assessing the model's accuracy and maximizing its performance by finding the best-suited clustering threshold for the model to give optimal results. The parameter tuning can be done either manually or automatically using methods such as Bayesian Optimization, Random Search, and Tree-structured Parzen Estimator (TPE). Also, the parameter tuning remains relative to how precise we want to be on finding or dropping matches while clustering, as there is always this trade-off between precision and recall.

3) *Serving*: This last step uses the trained and optimized model to identify matches and classify the records as "duplicate" or "not duplicate". Finally, the model returns clusters of partners. As duplication is transitive, clustering is performed on the matching pairs, so the same cluster's records are considered duplicates.

### D. Duplicates Cleaning

Once matches are gathered into clusters, data should be consolidated from many records into one. Many data fusion strategies could be used at this stage according to the strategic priorities of the data team (see Fig. 4). For example, if the process is more oriented towards data accuracy, the record of the most reliable source will likely be kept. Otherwise, the complete record will be held if the goal is to gather as much data as possible. Another data fusion strategy is to create a new record by merging the existing ones. In this case, a conflict resolution approach should be implemented to integrate duplicated columns. Multiple data fusion strategies were discussed in [28] [29].

### E. Continual Learning (Model Retraining)

Because of big data variability, data keeps changing constantly. Data could be changed regarding schema, statistical distribution, data quality, etc. This kind of change is known as data drift. In addition, data could also be exposed to a concept drift when the statistical properties of the target variable change over time [30]. Thus, the model's predictive performance may degrade over time because of data drift and concept drift. Therefore, it is crucial to adapt the model to data changes to ensure that the model accuracy is always maintained. For this, the model should be retrained after deployment according to an ML strategy called Continual Learning. Continual Learning is a process that automatically and continuously retrains a ML model with new data, which makes the model auto-adaptive and improves its performance. A critical use case of continual learning is recommendation systems that should always be updated with new data as user behavior changes over time. There are two approaches to performing continual learning:

- *Offline Mode (Batch learning)*: In this approach, the model is retrained from time to time with the new accumulated data.
- *Online Mode (Incremental Learning)*: the model is retrained sequentially with a live data stream.

With Online Continual Learning, the model does not decay following a data or concept drift as it is dynamically updated with new data patterns. The online mode is also a time effective solution as there is no need to store and manage large batches of accumulated data. On the other hand, the input data should constantly be monitored if the model is fed with insufficient data, the performance will be impacted instantly. The online mode remains suitable, especially in big data environments and real-time applications. Research has recently been conducted on Online Continual Learning, especially in a deep learning community. In [31], the authors have shown that algorithms and the architecture of neural networks impact continual learning performance. In [32], the authors have suggested a supervised training method for continual learning. The method's effectiveness was proven in three systems for continual online learning. In [33], the authors have introduced a new memory population approach (CBRS) for continual online learning that deals with imbalanced and temporally correlated data. Other pertinent methods for enhancing Online Continual Learning were suggested in [34] [35] [36]. For data deduplication, even if the deduplication model is trained with high-quality pairs, features defining duplications may change over time, especially when data is human input. Thus, new duplication features may come into play. Also, the used parts may become misleading, so they must be excluded or reweighted. Deduplication models are susceptible to duplication features, so a small features drift may drastically impact the model performance. In this regard, we suggest an Online Continual Learning approach for deduplication that consists of the following steps:

1) Building a dataset composed of new data and the found pairs during the serving.

- 2) Comparing and computing a similarity score of the built dataset.
- 3) Selecting the most similar pairs using a Threshold
- 4) Labeling the selected pairs as duplicates
- 5) Retraining the model with the new labeled pairs
- 6) Evaluating the model performance

This approach (Fig. 5) is executed in online mode, which makes it memory and time efficient, and hence, suitable for large datasets. This approach has also shown remarkable results in improving the model's accuracy. The model is continuously trained with new pairs, which allows updating the indexing rules with more pertinent ones. More details about the obtained results are provided in the next section.

In this section, we have presented the different steps of an end-to-end deduplication framework, including the data pre-processing, the labeling, the training, the serving phase, and finally, the retraining phase according to an online continual learning approach. For each phase, we have presented the different implementing techniques that could be used. Thus, the suggested framework is comprehensive and may be implemented differently depending on the intended use. Fig. 6 shows the machine learning pipeline of the whole framework. In the next section, we present how each step of the framework is implemented and the dataset and tools used for the implementation. Also, the suggested framework is compared against the existing approaches in terms of accuracy and scalability as the framework is designed to work in big data environments. Finally, a discussion is conducted about the possible evolutions.

## VI. IMPLEMENTATION

### A. Datasets Description

This section presents the implementation of the deduplication framework described in the previous section. The suggested framework was applied to 3 datasets:

**Dataset 1:** This first dataset is a built dataset with synthetic duplicated records. Indeed, to assess and evaluate the performance of the proposed strategy, the framework should be used for an extensive dataset with labeled duplicated records. Thus, we conducted research for datasets with two main criteria:

- A labeled dataset with a pre-defined state of true and false duplicates.
- Large Scale dataset with over 1M records.

Unfortunately, among the found datasets, no dataset matches the above criteria and, thus, was not appropriate for our use case. Indeed, previous research has also faced the same challenge as labeling big data sets manually are a very tedious and effortful task. To overcome this challenge, we built a labeled dataset with synthetic duplicated records using the Duplicate Generator tool DupGen [20] which allows for generating a synthetic dataset according to multiple criteria, such as the percentage of generated records and the changes made to data values. The built dataset contains over 1M records. It matches the Big Data characteristics not only in terms of Volume but also in terms of Variety, as the dataset

was gathered from multiple restaurant data sources with different formats and schemas. To ensure consistency, we have only kept standard fields: name, address, city, and type that refer to the restaurant's specialty. To stress our deduplication Framework, distinguishing features such as phone number and email were not considered even if they were available in all the datasets. The data sources used were clean of duplicates and were chosen from different countries so to avoid having common records between the datasets. After integrating and pre-processing source datasets, we have gathered a dataset with over 500 000 unique restaurants. The next step consists of creating duplicated records. For an accurate assessment, this process should not be done randomly. For this, we have reviewed restaurant datasets with real duplicates (these datasets were not suitable for our use due to their small volume) and tried to simulate duplicated data using the DupGen tool. Thus, we have noticed that most duplicated restaurants have either:

- Identical name, similar address and similar city and type
- Identical address, similar name, and similar city and type
- Similar name, similar address, and similar city and type

Also, we measured the average number of different characters between two duplicates for each column and applied the same distribution to our built dataset. Finally, we have created a dataset with over 1 M records with the following duplicates distribution: 80%: no duplication, 10%:1 duplication, 4%: 2 duplications, 2%: 3 duplications, 2%:4 duplications, 1%:5 duplications, and 1%: 6 to 10 duplications.

The number of duplicates was around 122 000, so the goal was to reduce over 1M records to about 878 000.

**Dataset 2:** The second dataset is a real companies name dataset containing 663000 records with 58700 duplicated records [37]. The dataset is prelabelled and intended for deduplication frameworks. This dataset was chosen to test our framework performance with a dataset of real-world values.

**Dataset 3:** The third dataset is a small dataset of 864 records with 112 duplicated records [38]. This dataset is prelabelled and was used by previous research to evaluate the deduplication methods. This dataset was chosen to compare our framework performance against the existing models.

Table I presents the characteristics of the three datasets used for our experiments. Before submitting the simulation results, we will first review the implementing tools and techniques in the next section.

TABLE I. DATASETS CHARACTERISTICS

Dataset	Records	Matchings	Threshold
Restaurant	864	112	0.76
Company	663000	58700	0.83
Built Dataset	1001300	122 000	0.75

### B. Adopted Tools and Techniques

The deduplication framework was developed on Apache Spark, suitable for Big Data. It was implemented in Python using Pyspark libraries such as Scikit-Learn for NLP and Fuzzy matching, Pandas, Scipy, and Numpy. For data pre-



processing, string functions were used as well as some Python's preprocessing packages such as NLTK, Stopwords, Unicode, Geocoder, LabelEncoder, RE (regular expressions) and Text blob. Then, data were first indexed using **Sorted neighborhood to build the training dataset**. The sorted neighborhood is an indexing technique that consists of sorting data values using the blocking key value and then moving a window of a fixed number of records over the sorted values. The sorted neighborhood index method is great when there is a relatively large amount of unstructured data. A recent study [14] has compared the indexing techniques for scalable linkage. It has shown that a sorted neighborhood is the most appropriate indexing method for big data in terms of execution time and accuracy. For more precision, we have applied weighted indexing to the restaurant dataset using the weights presented in Table II. With this parametrization, we suppose that the name and address are the most important columns to consider for restaurant's deduplication. These weights have allowed detecting accurately most of the pairs. No weights were applied to the company datasets with only one column (Company Name). Then, a similarity score is measured using **Cosine Similarity**. As mentioned before, various methods, such as Euclidean distance, and the Jaccard coefficient, can be used. However, Cosine Similarity is the most suited to measure text similarity, according to several studies [39] [40]. For more accuracy, the pairs are filtered out based on a min and max threshold range to keep only the most similar records. The selected records are then gathered into clusters to be used as a training dataset for the deduplication process. As mentioned before, the next step consists of dedupling the dataset using a ML algorithm. As mentioned before, in this approach duplicate detection is not based on a text comparison but on deduplication predicates and indexing rules generated by the model after the training. For this, we have used **Dedupe**. Dedupe is a Python library for accurate and scalable data deduplication and fuzzy matching based on ML [41] [42]. The first step consists of creating a dedupe instance for the dataset. Then, the dedupe instance is trained using the dataset built in the previous step. After the training phases, the model generates the indexing rules that will be used to detect similar records. In our case, one of the generated predicates was: (CommonTwoTokens, name), (SameSevenCharStart, name), (CommonThreeTokens, address). This means that the records with Names with the same two tokens AND Addresses with the same three tokens are considered duplicates. Once trained, the model can be used for deduplication using a semi-supervised clustering method, so similar records are clustered based on the provided labeled dataset. Then, the clustering threshold is tuned to get an optimized accuracy. Finally, duplicates are cleaned.

The Continual learning is performed in an online mode as it is carried out sequentially after each deduplication and is executed in real-time according to an automated machine learning pipeline. For this, a new dataset is built based on the

found pairs and additional 100 000 records. Then, the similarity of the detected pairs is evaluated using Cosine Similarity. A similarity threshold is set to select only the most matching pairs which are then labeled appropriately. The deduper is then retrained with the selected pairs to enhance the model's performance. An accuracy assessment is performed to evaluate the impact of continual learning on the model.

TABLE II. RESTAURANT DATASET INDEXING WEIGHTS

Name	Address	City	Type
0.55	0.35	0.05	0.05

### C. Results

#### a) Accuracy:

The framework was first applied to an extensive dataset of over 1M records with 122 000 duplicated records. 70% of the dataset was set to build the training dataset. Thus, the dataset was first indexed into pairs. Only 420 000 records were indexed as pairs. Then a similarity score was computed for the indexed pairs and a threshold of 0.75 was set to filter out only the most similar pairs. This first process resulted in 165000 pairs considered duplicates. The selected pairs were then gathered into clusters to detect records with more than one duplicate and were exported to a .csv file as a training dataset. Then, for each dataset, a dedupe model was trained using the built training dataset, tested, and optimized using the appropriate threshold. The performance of the framework was evaluated using the confusion matrix defined by the following metrics:

$$Precision = \frac{TP}{TP+FP} \quad (1)$$

$$Recall = \frac{TP}{TP+FN} \quad (2)$$

$$F - score = 2 * \frac{Precision*Recall}{Precision+Recall} \quad (3)$$

TP, FP, and FN are True Positive, False Positive, and False Negative, respectively.

The metrics above were measured for three different datasets: Restaurants, Companies, and our Built Big Dataset coming out with the results presented in Table III.

TABLE III. DEDUPLICATION FRAMEWORK EVALUATION

Dataset	Precision (%)	Recall (%)	F-s (%)
Restaurant	98,25%	100,00%	99,12%
Company	94,17%	98,13%	96,11%
Built Dataset	98,24%	96,48%	98,21%

It is worth noting that the framework accuracy has evolved considerably after applying online continual learning. For our built dataset, the framework detected 117700 out of 122 000 duplicated records with an F-score of 98,21%. Indeed, the resulting F-score was initially 97,05% and has increased by 1,16% after applying the continual learning process to the model with an additional dataset of 100 000 records.

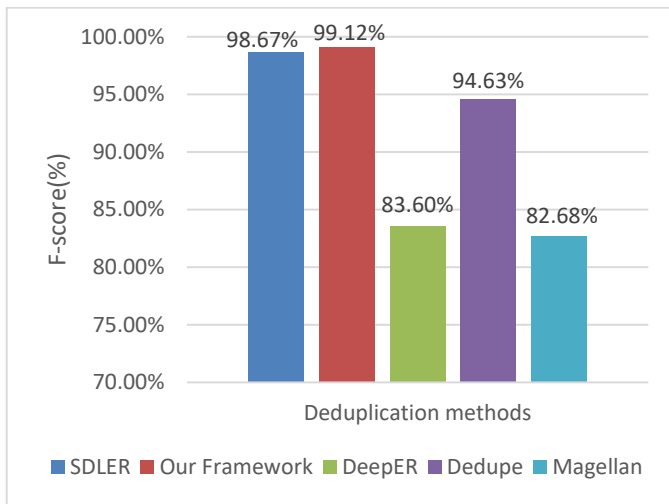


Fig. 7. F-Score Comparison.

As mentioned previously, we chose the third dataset to compare our framework’s performance against the existing models that have used the same dataset (restaurant dataset), such as SDLER [18], DeepER [43], Magellan [44], and Dedupe [42]. Fig. 7 compares the F-score achieved by each model when applied to the Restaurant dataset. The obtained results show that the proposed framework provides the best results in terms of accuracy. When dealing with big data, the execution time is yet another factor that should be considered besides accuracy. We present in the next part the time complexity of the proposed framework.

#### b) Scalability

As the framework is intended to be used in a big data environment, the framework scalability also needs to be ensured. Table IV shows the processing time and the corresponding dataset size. Thus, the framework has shown acceptable results in terms of processing time with a linear complexity  $O(n)$ . Indeed, the framework is based on scalable methods such as Sorted Neighborhood, Cosine Similarity, and Dedupe having a linear complexity and hence, are suitable for Big Data [45] [41].

TABLE IV. PROCESSING TIME

Dataset	Records	Processing Time
Restaurant	864	~3 m
Company	663000	~ 1h
Built Dataset	1001300	~ 3h30

#### c) Framework Limitations

A second phase of the implementation consists of scrambling the built dataset intentionally by feeding the datasets with more challenging duplicates. The goal is to uncover the framework limitations and discover how the accuracy is impacted by the inferior and very poor data quality and to what extent the framework remains suitable for use. For this, we have unfiltered in the dataset extreme cases of non-duplicates where for example the name and the address are similar, but the records are not duplicates. The framework was applied to a very poor big data quality to uncover the limitations of the framework. The dataset was then scrambled

progressively with a very poor-quality dataset, and the accuracy was assessed in each round. The F-score has decreased in each round, as shown in Table V. Thus, it turns out that the framework resists and remains functional in a half-scrambled dataset with an F-score of 88,2%. Therefore, the accuracy is acceptably impacted by a very-poor biasing dataset.

TABLE V. F-SCORE EVOLUTION IN A VERY POOR-QUALITY DATASET

Percentage of scrambled data	F-s (%)
20 % of the very poor-quality dataset	97,9%
35 % of the very poor-quality dataset	94,8%
50 % of the very poor-quality dataset	88,2%
60 % of the very poor-quality dataset	83,4%

#### D. Discussion

Although significant efforts have been made in recent years for data deduplication, there are still challenges to be addressed, especially for big data. Indeed, data uniqueness as a quality metric depends highly on other quality metrics such as completeness, accuracy, validity, etc. For example, even if we have imputed data during the pre-processing phase, most imputation methods are not accurate, which can impact the deduplication accuracy as data is credited with inaccurate values. Meanwhile, ignoring missing values will negatively affect the model accuracy, especially in a big data environment where most of the data are incomplete. On the other hand, deduplication can also impact the other metrics, as the cleaning phase consists of keeping the most accurate, complete, or recent record. In some cases, records can even be merged. All these changes have a high impact on the other metrics. Thus, data deduplication could not be improved separately and, therefore, should be addressed in a more comprehensive approach that considers this strong relationship between the quality metrics. Continual Learning is yet another research area that needs more focus. Even if Continual Learning has been around for more than 20 years, there are challenges that still need to be addressed, such as catastrophic forgetting, auditing, mentoring, evaluating continual learning techniques, etc. In addition to these challenges, new issues have been raised with big data, such as handling memories, learning for streaming multimodal data, model saturation, etc. Thus, continual learning is not already in its explosion, and further research is needed. However, it is safe to say that Continual Learning will become increasingly crucial as ML models could not be effectively performed without accumulating the learned knowledge.

#### VII. CONCLUSION

While data deduplication has been the subject of several studies in the last decade, some challenges remain, especially in the Big Data Era. In this article, we have reviewed the most recent big data deduplication frameworks suggested in the literature. We also proposed a novel end-to-end big data deduplication framework based on a Semi-supervised clustering approach. The experiments have shown that the framework outperforms the existing big data deduplication approaches with an F-score of 98,21%. The suggested framework is also extended with an online continual learning phase that continuously improves the deduplication model

performance and increases the model accuracy by 1,16%. In future work, we aim to enhance our framework by reducing the error rate when used on a very-poor quality dataset. Also, we aim to extend our framework to address more quality dimensions.

#### REFERENCES

- [1] Y. Gahi, M. Guennoun, and H. T. Mouftah, 'Big Data Analytics: Security and privacy challenges', in 2016 IEEE Symposium on Computers and Communication (ISCC), Jun. 2016, pp. 952–957. doi: 10.1109/ISCC.2016.7543859.
- [2] I. El Alaoui, Y. Gahi, and R. Messoussi, 'Big Data Quality Metrics for Sentiment Analysis Approaches', in Proceedings of the 2019 International Conference on Big Data Engineering, New York, NY, USA, Jun. 2019, pp. 36–43. doi: 10.1145/3341620.3341629.
- [3] I. E. Alaoui and Y. Gahi, 'The Impact of Big Data Quality on Sentiment Analysis Approaches', *Procedia Comput. Sci.*, vol. 160, pp. 803–810, Jan. 2019, doi: 10.1016/j.procs.2019.11.007.
- [4] W. Elouataoui, I. E. Alaoui, and Y. Gahi, 'Data Quality in the Era of Big Data: A Global Review', in *Big Data Intelligence for Smart Applications*, Y. Baddi, Y. Gahi, Y. Maleh, M. Alazab, and L. Tawalbeh, Eds. Cham: Springer International Publishing, 2022, pp. 1–25. doi: 10.1007/978-3-030-87954-9\_1.
- [5] I. E. Alaoui, Y. Gahi, and R. Messoussi, 'Full Consideration of Big Data Characteristics in Sentiment Analysis Context', in 2019 IEEE 4th International Conference on Cloud Computing and Big Data Analysis (ICCCBDA), Apr. 2019, pp. 126–130. doi: 10.1109/ICCCBDA.2019.8725728.
- [6] Elouataoui, W.; El Alaoui, I. and Gahi, Y. (2022). Metadata Quality Dimensions for Big Data Use Cases. In Proceedings of the 2nd International Conference on Big Data, Modelling and Machine Learning - BML, ISBN 978-989-758-559-3, pages 488-495. DOI: 10.5220/0010737400003101
- [7] V. Christophides, V. Efthymiou, T. Palpanas, G. Papadakis, and K. Stefanidis, 'An Overview of End-to-End Entity Resolution for Big Data', *ACM Comput. Surv.*, vol. 53, no. 6, p. 127:1-127:42, Dec. 2020, doi: 10.1145/3418896.
- [8] D. Tranfield, D. Denyer, and P. Smart, 'Towards a Methodology for Developing Evidence-Informed Management Knowledge by Means of Systematic Review', *Br. J. Manag.*, vol. 14, no. 3, pp. 207–222, 2003, doi: 10.1111/1467-8551.00375.
- [9] O. Azeroual, M. Jha, A. Nikiforova, K. Sha, M. Alsmirat, and S. Jha, 'A Record Linkage-Based Data Deduplication Framework with DataCleaner Extension', *Multimodal Technol. Interact.*, vol. 6, no. 4, Art. no. 4, Apr. 2022, doi: 10.3390/mti6040027.
- [10] Simonini, Giovanni; Saccani, Henrique; Gagliardelli, Luca; Zecchini, Luca; Benevento, Domenico; Bergamaschi, Sonia. The Case for Multi-task Active Learning Entity Resolution / (2021).
- [11] M. Pikies and J. Ali, 'Analysis and safety engineering of fuzzy string matching algorithms', *ISA Trans.*, vol. 113, pp. 1–8, Jul. 2021, doi: 10.1016/j.isatra.2020.10.014.
- [12] Y. Gahi and I. El Alaoui, 'Machine Learning and Deep Learning Models for Big Data Issues', in *Machine Intelligence and Big Data Analytics for Cybersecurity Applications*, Y. Maleh, M. Shojafar, M. Alazab, and Y. Baddi, Eds. Cham: Springer International Publishing, 2021, pp. 29–49. doi: 10.1007/978-3-030-57024-8\_2.
- [13] I. El Alaoui, Y. Gahi, R. Messoussi, A. Todoskoff, and A. Kobi, 'Big Data Analytics: A Comparison of Tools and Applications', in *Innovations in Smart Cities and Applications*, Cham, 2018, pp. 587–601. doi: 10.1007/978-3-319-74500-8\_54.
- [14] S. YEDDULA and K. LAKSHMAIAH, 'INVESTIGATION OF TECHNIQUES FOR EFFICIENT & ACCURATE INDEXING FOR SCALABLE RECORD LINKAGE & DEDUPLICATION', *Int. J. Comput. Commun. Technol.*, vol. 6, no. 1, Sep. 2020, doi: 10.47893/IJCCCT.2015.1275.
- [15] X. Chen, 'Towards Efficient and Effective Entity Resolution for High-Volume and Variable Data', p. 167, 2020, doi: 10.25673/35204
- [16] El-Ghafar, R.M., El-Bastawissy, A.H., Nasr, E.S., Gheith, M.H., & Independent Researcher, C.E. (2021). An Effective Entity Resolution Approach for Big Data. *International Journal of Innovative Technology and Exploring Engineering*.
- [17] V. Efthymiou, K. Stefanidis, and V. Christophides, 'Big data entity resolution: From highly to somehow similar entity descriptions in the Web', in 2015 IEEE International Conference on Big Data (Big Data), Oct. 2015, pp. 401–410. doi: 10.1109/BigData.2015.7363781.
- [18] A. Nguetilbaye, H. Wang, D. A. Mahamat, and I. A. Elgendy, 'SDLER: stacked dedupe learning for entity resolution in big data era', *J. Supercomput.*, vol. 77, no. 10, pp. 10959–10983, Oct. 2021, doi: 10.1007/s11227-021-03710-x.
- [19] T. Papenbrock, A. Heise, and F. Naumann, 'Progressive Duplicate Detection', *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 5, pp. 1316–1329, May 2015, doi: 10.1109/TKDE.2014.2359666.
- [20] El-Ghafar, R. M. A., El-Bastawissy, A. H., Nasr, E. S., & Gheith, M. H. (2020). An Efficient Multi-Phase Blocking Strategy for Entity Resolution in Big Data. In *International Journal of Innovative Technology and Exploring Engineering* (Vol. 9, Issue 9, pp. 254–263).
- [21] W. Elouataoui, I. El Alaoui, and Y. Gahi, 'Metadata Quality in the Era of Big Data and Unstructured Content', in *Advances in Information, Communication and Cybersecurity*, Cham, 2022, pp. 110–121. doi: 10.1007/978-3-030-91738-8\_11.
- [22] A. Z. Faroukhi, I. El Alaoui, Y. Gahi, and A. Amine, 'Big data monetization throughout Big Data Value Chain: a comprehensive review', *J. Big Data*, vol. 7, no. 1, p. 3, Jan. 2020, doi: 10.1186/s40537-019-0281-5.
- [23] Experian.com. (n.d.). Retrieved August 12, 2022, from <http://experian.com/assets/decision-analytics/white-papers/the%20state%20of%20data%20quality.pdf>
- [24] Chaitra. (2021, June 22). Understanding data deduplication - and why it's critical for moving data to the cloud. Druva. Retrieved August 12, 2022, from <https://www.druva.com/blog/a-simple-definition-what-is-data-deduplication>
- [25] Druva inc. (n.d.). Customers win with Druva and AWS. Druva. Retrieved August 12, 2022, from <https://content.druva.com/cb-customers-win-with-druva-aws?x=4if2hg>
- [26] A. Z. Faroukhi, I. El Alaoui, Y. Gahi, and A. Amine, 'An Adaptable Big Data Value Chain Framework for End-to-End Big Data Monetization', *Big Data Cogn. Comput.*, vol. 4, no. 4, Art. no. 4, Dec. 2020, doi: 10.3390/bdcc4040034.
- [27] Buschi, N. (2021, June 23). Top 5 challenges making data labeling ineffective. Dataloop. Retrieved June 4, 2022, from <https://dataloop.ai/blog/data-labeling-challenges/>.
- [28] Christen, P.: Further topics and research directions. In: Christen, P. (ed.) *Data Matching*, pp. 209–228. Springer, Heidelberg (2012)
- [29] D. Elkington, X. Zeng, and R. Morris, 'Resolving and merging duplicate records using machine learning', *US20160357790A1*, Dec. 08, 2016.
- [30] Komolafe, A. (2022, July 22). Retraining model during deployment: Continuous training and continuous testing. neptune.ai. Retrieved August 25, 2022, from <https://neptune.ai/blog/retraining-model-during-deployment-continuous-training-continuous-testing>
- [31] S. I. Mirzadeh et al., 'Architecture Matters in Continual Learning', *arXiv*, arXiv:2202.00275, Feb. 2022. doi: 10.48550/arXiv.2202.00275.
- [32] J. Gallardo, T. L. Hayes, and C. Kanan, 'Self-Supervised Training Enhances Online Continual Learning', *arXiv*, arXiv:2103.14010, Oct. 2021. doi: 10.48550/arXiv.2103.14010.
- [33] A. Chrysakos and M.-F. Moens, 'Online Continual Learning from Imbalanced Data', in Proceedings of the 37th International Conference on Machine Learning, Nov. 2020, pp. 1952–1961.
- [34] E. Fini, S. Lathuilière, E. Sanginetto, M. Nabi, and E. Ricci, 'Online Continual Learning under Extreme Memory Constraints', *arXiv*, arXiv:2008.01510, Jan. 2022. doi: 10.48550/arXiv.2008.01510.
- [35] H. Koh, D. Kim, J.-W. Ha, and J. Choi, 'Online Continual Learning on Class Incremental Blurry Task Configuration with Anytime Inference', presented at the International Conference on Learning Representations, Sep. 2021.

- [36] C. Wiwatcharakoses and D. Berrar, 'A self-organizing incremental neural network for continual supervised learning', *Expert Syst. Appl.*, vol. 185, p. 115662, Dec. 2021, doi: 10.1016/j.eswa.2021.115662.
- [37] Caesarlupum. (2019, November 19). Deduping & Record Linkage. Kaggle. Retrieved August 12, 2022, from <https://www.kaggle.com/code/caesarlupum/deduping-record-linkage>
- [38] Duplicate detection, record linkage, and identity uncertainty: Datasets. (n.d.). Retrieved August 12, 2022, from <https://www.cs.utexas.edu/users/ml/riddle/data.html>
- [39] Cosine similarity. Cosine Similarity - an overview | ScienceDirect Topics. (n.d.). Retrieved August 25, 2022, from <https://www.sciencedirect.com/topics/computer-science/cosine-similarity>
- [40] M. Kirişci, cosine similarity FFS. 2022.
- [41] F. Gregg, dedupe: A python library for accurate and scaleable data deduplication and entity-resolution. Accessed: Jun. 18, 2022. Available: <https://github.com/dedupeio/dedupe>.
- [42] Vintasoftware. Deduplication-slides/slides.ipynb at vintasoftware/deduplication-slides. GitHub. Retrieved August 12, 2022, from <https://github.com/vintasoftware/deduplication-slides/blob/631389413a558ea83a407a47870253325b7b068e/slides.ipynb>
- [43] M. Ebraheem, S. Thirumuruganathan, S. Joty, M. Ouzzani, and N. Tang, 'Distributed representations of tuples for entity resolution', *Proc. VLDB Endow.*, vol. 11, no. 11, pp. 1454–1467, Jul. 2018, doi: 10.14778/3236187.3236198.
- [44] A. Doan et al., 'Human-in-the-Loop Challenges for Entity Matching: A Midterm Report', in *Proceedings of the 2nd Workshop on Human-In-the-Loop Data Analytics*, New York, NY, USA, May 2017, pp. 1–6. doi: 10.1145/3077257.3077268.
- [45] Project C CSE 494/598 Hemal Khatri - Arizona State University. (n.d.). Retrieved August 12, 2022, from [https://rakaposhi.eas.asu.edu/cse494/f05-projects/ProjC\\_Hemal.pdf](https://rakaposhi.eas.asu.edu/cse494/f05-projects/ProjC_Hemal.pdf)

# Student's Performance Prediction based on Personality Traits and Intelligence Quotient using Machine Learning

Samar El-Keiey<sup>1</sup>, Dina ElMenshawy<sup>2</sup>, Ehab Hassanein<sup>3</sup>  
Information Systems Department  
Faculty of Computers and Artificial Intelligence  
Cairo University  
Egypt

**Abstract**—Apparently, most life activities that people perform depend on their unique characteristics. Personal characteristics vary across people, so they perform tasks in different ways based on their skills. People have different mental, psychological, and behavioral features that affect most life activities. This is the same case with students at various educational levels. Students have different features that affect their academic performance. The academic score is the main indicator of the student's performance. However, other factors such as personality features, intelligence level, and basic personal data can have a great influence on the student's performance. This means that the academic score is not the only indicator that can be used in predicting students' performance. Consequently, an approach based on personal data, personality features, and intelligence quotient is proposed to predict the performance of university undergraduates. Five machine learning techniques were used in the proposed approach. In order to evaluate the performance of the proposed approach, a real student's dataset was used, and various performance measures were computed. Several experiments were performed to determine the impact of various features on the student's performance. The proposed approach gave promising results when tested on the dataset.

**Keywords**—Prediction; student performance; machine learning; personality; intelligence quotient

## I. INTRODUCTION

People have a wide range of cognitive abilities, including intelligence, memory, attention, and so on. People can carry out brain-based operations or activities in a variety of ways and for varying lengths of time. Two people can perform the same operation in two different ways. This is because people's personal characteristics differ, and most life activities are dependent on these personal characteristics.

Each person has certain abilities which he/she uses to deal with various real-life activities. Also, people have different ways to process and memorize information. Actually, all operations or activities a person performs are based on his/her features either the behavioral features or the psychological ones.

For example, the tasks that need some sort of intelligence, people who have higher intelligence quotients usually finish their assigned tasks in fewer steps and in less time than others who have lower intelligence quotients. Moreover, people who

have higher memory skills usually remember things better and more quickly than other people who have lower memory skills.

Apparently, most life operations or actions which people perform are done based on their various skills. The same idea goes for students at all educational levels. Students have different characteristics either psychologically or mentally which affect their academic performance.

Some students perform better in theoretical questions that need remembering skills while other students perform better in practical questions which need brain-based skills. Mainly the academic score of a student depends on his/her unique characteristics.

The main purpose of this research is to measure the impact of the personality traits and the intelligence of students on their academic performance. In addition to these features, some personal data and the academic score have been used to predict the academic performance of undergraduate students.

To the best of our knowledge, no work exists answering the question of the impact of personal data, personality traits, intelligence quotient, and academic score on the student's performance.

The research questions of this research are:

- 1) What effect does the Intelligence Quotient (IQ) have on the undergraduate student's performance?
- 2) What effect do personality traits have on the undergraduate student's performance?
- 3) What effect does the combination of IQ, personality traits, personal data, and academic score have on the undergraduate student's performance?
- 4) What are the most significant features in the student's performance prediction?

## II. STUDENT PERFORMANCE PREDICTION

Student performance prediction has played a significant role in educational systems. Predicting student performance helps students to select appropriate courses which match their skills as the student's performance can vary across different courses. Prediction helps students choose courses that match

their abilities. Also, student performance prediction can assist in designing appropriate future study plans for students. Beyond predicting student performance, it helps teachers and managers to monitor and support the students, and to offer training programs to achieve the best results. Other benefits of student performance prediction are reducing official warning signs and discarding students for inefficiency.

Machine learning techniques have played a significant role in the creation of effective educational systems over the past two decades. These techniques helped in offering better learning techniques and in enhancing the academic performance of students [1].

Applying machine learning techniques in educational systems has played a crucial role in discovering concealed and unexpected methods to impart knowledge across all educational levels. As a result, some prediction models have been proposed by numerous researchers to enhance the student's performance and learning quality as in [2], [3].

### III. MOTIVATION

Student performance mainly can be predicted through historical records of quizzes and exams, and Grade Point Average (GPA). This is usually common in various educational levels and ages. However, sometimes there are other factors that may influence a student's performance such as mentality skills, behavioral characteristics, cognitive skills, personality traits, and psychological factors.

Also, some of these factors may affect the performance of students at certain ages but may have no effect at other ages or at other educational levels. For example, personality traits may have a great influence on older students like university students but may not have a noticeable effect on younger students at primary or high school levels.

Moreover, the factors that are considered important for student performance prediction for preschool students may not have the same importance as for older students like undergraduates in universities. Also, gender can play a role in student performance prediction. For example, sometimes male students have better scores in different courses or specialties than female students and vice versa.

The importance of a certain factor or a feature in student performance prediction usually depends on the age or the educational level in which the student is enrolled. Consequently, in this research, we focus on a certain educational level which is undergraduates in universities.

As a result, in this paper, an approach based on personal data, personality traits, and intelligence quotient is proposed to predict the student's performance of university undergraduate students.

The main contributions of this paper are as follows:

1) Proposed an approach that utilized academic score, gender, region/city, number of brothers/sisters, Intelligence Quotient (IQ), and personality features to predict student's performance.

2) Applied five machine learning techniques in the proposed approach to predict student's performance.

3) Applied the proposed approach on real students' datasets.

4) Compared the performance of different features across the five machine learning techniques.

5) Predicted the best indicators that assist in student performance prediction.

The rest of the paper is organized as follows: Section 4 presents the literature work on students' performance prediction. Section 5 presents the proposed approach. Section 6 presents the results. Section 7 presents the analysis and discussion of the results. Section 8 presents the conclusion and future work.

### IV. RELATED WORK

In the following paragraphs, related work on students' performance prediction using different techniques is presented.

In recent decades, many attempts have been made to predict students' academic performance before students start the learning process to make their outcomes predictable. This is also necessary for instructors to know the areas where students have defects so that students' skills in these areas can be improved. By predicting future results in a timely manner, instructors can know the areas that need improvement while teaching students.

Educational institutions and governments also want to know the performance of the current educational system in order to perform improvements in the long term. In [4], the authors showed that students' performance depends on various factors such as demographics, behavior, previous outcomes, and habits. Unexpected factors, such as the address of students, had a great impact on the student's performance.

Many studies have been conducted to discover the effect of various variables on student academic performance. These factors are not the same all over the world and may vary from university to university, from university to school, and also from individual to individual.

In today's world, data has become very powerful, and machine learning can be very helpful in harnessing the power of this data. Machine learning techniques, along with deep learning techniques, have played a very important role in predicting student academic performance.

Various machine learning and deep learning techniques, such as Support Vector Machine (SVM), Neural Networks (NN), and clustering have been studied on different datasets in different institutions to find hidden and unexpected patterns. Several machine learning techniques were applied in [5].

In [6], various models for academic performance prediction have been developed using Decision Tree (DT), Naïve Bayes (NB), and Rule-Based (RB) for the Bachelor of Computer Science students at the University Sultan Zainal Abidin. The results showed that DT and RB provided better accuracies than NB.



In [7], the authors applied SVM, NN, DT, and NB models on two independent datasets. The results showed that SVM performed better than the other methods in terms of statistical significance.

Predicting student performance is a key challenge in the educational process that uses technology to help students towards success [8]. Previous research has detected several factors that influence the student's performance. Some of these factors are student demographics such as gender [9], previous academic grades [10], and interaction with the learning environment [11], [12].

In [13], DT, NB, logistic regression, SVM, K-nearest neighbors, iterative minimum optimization, and NN were utilized to study the student's performance in final undergraduate exams. Logistic regression performed the best compared to other models used in this study, with an accuracy of 66%.

In [14], the proposed model applied NN, logistic regression, and SVM algorithms in a virtual school learning environment.

In [15], recurrent neural networks (RNNs) for detailed knowledge and engagement studies were used. In this study, the authors achieved an accuracy value of 88.3%. SVM, linear discriminant analysis, random forests, K-nearest neighbors, and classification regression trees (CART) were used in this study. Random forests performed the best, with an accuracy of 90%.

In [16], the research applied a logistic regression prediction model on some variables which are level of involvement, level of prestige, level of visibility, number of student visits, and management system by subject, experience, age, and gender. The predictive models used in the analysis were NN, DT, and NB using bagging, boosting, and ensemble techniques. The DT classifier gave the highest accuracy value of 82%.

In [17], the authors provided the most comprehensive assessment to ensure the strength of the relationship between the big five personality traits and academic performance through the synthesis of 267 independent samples (N = 413,074) in 228 original studies. Also, the progressive validity of personality traits beyond cognitive ability in predicting academic performance was examined.

In [18], the research introduced a Contextual Cognitive Skill Scores (CCSS) approach to predict the student's performance. To determine the CCSS score, the cognitive skills required to solve the exam questions were combined with the exam scores. In doing so, the authors focused on generalizing the student's performance to understand its utility in predicting student risk profiles.

Exam questions and their scores vary by course and by exam. Such data heterogeneity is very important when generalizing the model. This is because feature variation affects the results. Therefore, the authors built the CCSS so that the feature space is the same for all courses so that students' performance can be visualized in the same dimension.

In [19], the proposed model predicted the success of students based on their behavioral patterns/activities in the learning process. Six machine learning prediction models were presented. Then, accuracy measures were computed to evaluate the proposed model.

In [20], data was collected from different students with different parameters, then these parameters were analyzed using techniques similar to those used in [21]. After that, decision trees for all considered attributes were built. Next, "if-then" rules for the relationship between various attributes and the student's performance were generated.

In [22], some machine learning algorithms were utilized to predict and classify different types of educational data. Three machine learning algorithms, namely, backpropagation (BP), support vector regression (SVR), and long short-term memory (LSTM) were used to predict the student's performance. In addition to these algorithms, a Gradient Boosting Classifier (GBC) is implemented in the classification phase.

## V. PROPOSED APPROACH

In this section, the proposed approach is presented along with the features used, the techniques applied, and the dataset used.

The main idea of our proposed approach is to discover the impact of the big five personality traits and the intelligence quotient on the performance of students when these features are integrated with the academic score feature and some personal data namely, gender, number of brothers/sisters and the region/city where each student lives in. The proposed approach was implemented in Python.

### A. Features Used and Data Collection

For the proposed approach, the data was collected from students in the Faculty of Computers and Artificial Intelligence, Cairo University. The data was collected from students from various levels, namely, 1<sup>st</sup>, 2<sup>nd</sup>, and 3<sup>rd</sup> year undergraduate students. Approximately 300 students shared their data and their grades in different courses which are an introduction to database systems, fundamentals of computer science, and Web-based information systems.

The features collected are:

- Gender
- Number of brothers/sisters
- Region/city
- GPA
- IQ
- Five personality traits, namely,
  1. Openness (O)
  2. Conscientiousness (C)
  3. Extraversion (E)
  4. Agreeableness (A)
  5. Neuroticism (N)

The 300 students who shared their data were divided into 117 females and 183 males. The average GPA is 2.9 and the average IQ is 3.7.

A snapshot of the dataset is shown in Fig. 1.

GPA	Gender	brothers/sisters	O	C	E	A	N	IQ	city/region
2.32	Male	3	4	2	3	3	3	5	Helwan / Cairo
2.9	Male	1	2	5	1	4	1	4	Cairo
3.6	Male	2	3	5	5	1	1	4	Giza
2.1	Male	2	3	4	4	3	4	2	Helwan/Egypt
3.65	Female	2	2	5	2	5	2	5	Ciario
2.99	Male	2	2	5	4	5	2	4	Giza
2.9	Male	2	2	3	4	4	4	3	Maadi , Cairo
2.67	Male	3	1	4	3	3	1	4	Nasr city
3.4	Male	3	2	5	3	3	1	5	fisal-guza
3	Female	2	4	3	3	3	5	4	giza
3.45	Female	2	4	4	4	4	3	5	New cairo
3.67	Female	2	4	5	4	4	4	5	Maadi-Cairo
3.95	Male	2	3	5	1	4	1	5	Giza/Haram
2.93	Male	3	2	3	3	4	3	4	giza
3.7	Female	2	3	5	4	3	1	4	Cairo
2.96	Male	2	3	4	3	4	1	4	Giza
3.01	Female	2	4	3	4	3	4	4	Giza
1.91	Male	2	2	1	4	1	5	2	Cairo
2.7	Female	3	3	3	3	1	4	4	Giza
2.87	Male	2	2	3	2	3	3	4	Cairo
3.43	Female	2	3	4	3	3	2	5	cairo

Fig. 1. Snapshot of the Dataset.

### B. Data Preprocessing

The data was collected from a questionnaire via Google form so some preprocessing steps were performed so that the data can be used in the proposed prediction approach. The preprocessing steps focused mainly on removing noise and minimizing redundancy. Also, some features whose values were text were converted to numeric so that they can be used in the prediction model.

Moreover, the numeric scores were converted to categorical classes which are low, medium, and high so that the class label to which each student belongs can be predicted. Each categorical class included a range of values, as follows:

- Low (0 - <40)
- Medium (40 - <80)
- High (>= 80)

### C. Techniques Used

Mainly, five machine learning techniques were used in our proposed approach in predicting the student's performance. The used techniques are:

1) **k-nearest neighbor (KNN)**: KNN is a non-parametric supervised machine learning classifier algorithm, which is used for regression and classification [23]. After performing various experiments to determine the best value of k, a value of 6 gave the best accuracy, as presented in Fig. 2.

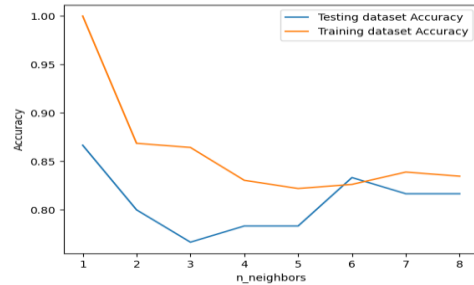


Fig. 2. The Best K Value.

2) **Support Vector Machine (SVM)** is a supervised learning model used for regression as well as for classification [24].

3) **Decision Tree (DT)** is a non-parametric supervised machine learning classifier algorithm, which is used for regression and classification. It has a hierarchical tree structure [25].

4) **Random Forest (RF)** is a classification algorithm used for classification and regression. It consists of several decision trees [26].

5) **Naive Bayes (NB)** is a probabilistic classifier that applies Bayes' theorem [27].

### D. Features Importance

To know the most influential features in the prediction, the Scikit-learn (Sklearn) python library and the feature importance python function [28] were used to determine the most important features as presented in Fig. 3.

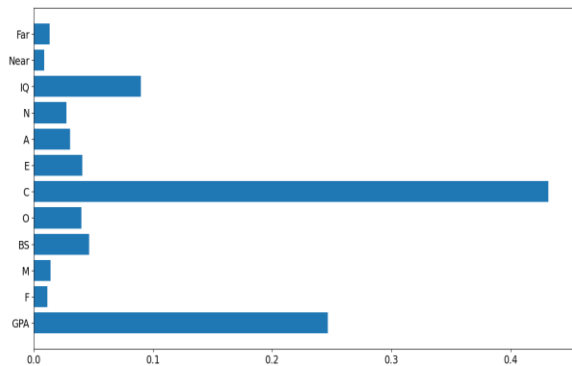


Fig. 3. The Most Important Features.

As shown in Fig. 3, the features are abbreviated as follows:

- GPA: Grade Point Average
- F: Female
- M: Male
- BS: Number of Brothers/Sisters
- O: Openness
- C: Conscientiousness
- E: Extraversion
- A: Agreeableness

- N: Neuroticism
- IQ: Intelligence Quotient

After performing several experiments, it was deduced that the most important features that have the highest significance on the student’s performance are Conscientiousness, GPA, and IQ.

Also, it was noticed that gender and location have the least significant on the student’s performance. The location feature is not of great importance, so our proposed approach can be applied easily in online learning.

E. Correlation Matrix

To know the correlation and the dependencies between the features, the correlation matrix was constructed, and it is shown in Fig. 4.

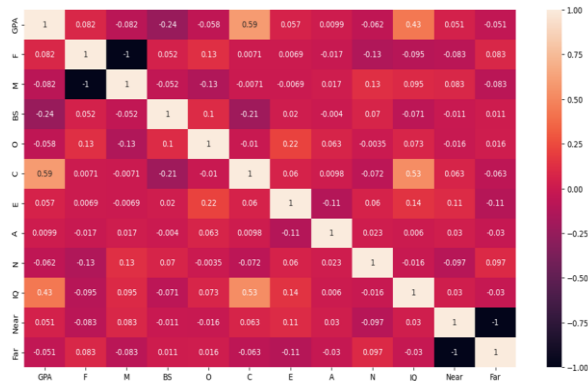


Fig. 4. Correlation Matrix.

The correlation between the features is the indicator of the extent of the effect of one feature on another one. The higher the value (either positively or negatively) between a pair of features, the higher is the correlation between them.

On the other hand, the less correlated features gave values closer to zero. When a value becomes closer to zero, this means that the features are less correlated.

When a value of a certain feature increases and the value of the other correlated feature increases, this denotes a positive correlation. On the other hand, when a value of a certain feature increases and the value of the other correlated feature decreases, and vice versa, this denotes a negative correlation.

After performing several experiments on the various features, the most correlated features were GPA, Conscientiousness, and IQ.

VI. RESULTS

The collected dataset was used to evaluate the performance of the proposed approach. The prediction accuracy, precision, recall, F1 measure, and the confusion matrix (true positive, true negative, false positive, and false negative) were computed for all techniques. Also, the results are compared across the five techniques. The results are presented in Table I.

TABLE I. PREDICTION RESULTS OF THE FIVE TECHNIQUES

Technique	Accuracy	Precision	Recall	F1 measure
KNN	0.85	0.84	0.85	0.84
NB	0.86	0.86	0.86	0.86
RF	0.83	0.82	0.83	0.83
DT	0.90	0.89	0.90	0.89
SVM	0.88	0.87	0.88	0.88

As shown in Table I, the decision tree technique gave the best performance among the other techniques with accuracy = 0.90%, precision = 0.89%, recall = 0.90% and F1 measure = 0.89%.

A. Confusion Matrix for All Techniques

The confusion matrix is used to analyze the performance of the classification techniques by computing the True positive (TP), the True Negative (TN), the False Positive (FP), and the False Negative (FN) for the testing data which are described as follows.

- TP: The true positive value is the case in which the actual value and expected value are identical [29].
- TN: A class’s True Negative value is the sum of all columns and rows, except those for the class for which we are computing the values [29].
- FP: A class’s False-positive value is the sum of all the values in the relevant column, except for the TP value [29].
- FN: A class’s False-negative value is the sum of the values in the relevant rows, except for the TP value [29].

Using the confusion matrix, we can assess the model’s performance in terms of recall, precision, and accuracy.

The confusion matrices for the prediction classification techniques are shown below in Tables II to VI.

1) K-Nearest Neighbor

TABLE II. KNN CONFUSION MATRIX

	High	Medium	Low
High	31	0	0
Medium	1	11	4
Low	2	2	9

Predicted values

Actual values

2) Naïve Bayes

TABLE III. NAIVE BAYES CONFUSION MATRIX

	<b>High</b>	<b>Medium</b>	<b>Low</b>
<b>High</b>	31	0	0
<b>Medium</b>	1	13	2
<b>Low</b>	3	2	8

Predicted values

3) Random Forest

TABLE IV. RANDOM FORESTS CONFUSION MATRIX

	<b>High</b>	<b>Medium</b>	<b>Low</b>
<b>High</b>	31	0	0
<b>Medium</b>	1	12	3
<b>Low</b>	3	2	8

Predicted values

4) Decision Tree

TABLE V. DECISION TREE CONFUSION MATRIX

	<b>High</b>	<b>Medium</b>	<b>Low</b>
<b>High</b>	31	0	0
<b>Medium</b>	0	14	2
<b>Low</b>	3	2	9

Predicted values

5) Support Vector Machine

TABLE VI. SUPPORT VECTOR MACHINE CONFUSION MATRIX

	<b>High</b>	<b>Medium</b>	<b>Low</b>
<b>High</b>	31	0	0
<b>Medium</b>	0	13	3
<b>Low</b>	2	2	9

Predicted values

VII. RESULT ANALYSIS AND DISCUSSION

The proposed approach proved that the big five personality traits, especially the “conscientiousness” feature, are the most significant features in predicting student’s performance for undergraduate students in the Faculty of Computers and Artificial Intelligence, Cairo University in Egypt.

The intelligence quotient score also has a significant role in our prediction approach.

In order to test the efficiency of our proposed approach in integrating several features, personality traits and IQ features were removed from the dataset and the results were compared before and after removing these features as will be described in the following paragraphs.

A. The Impact of Removing the Big Five Personality Traits

Another experiment has been conducted to evaluate the performance of all techniques after removing all big five personality traits features from the dataset. This experiment was performed to monitor the impact of the big five personality traits on the academic performance of Egyptian students. The results are presented in Table VII.

TABLE VII. PREDICTION RESULTS WITHOUT BIG FIVE PERSONALITY FEATURES

Technique	Accuracy	Precision	Recall	F1 measure
<b>KNN</b>	0.66	0.68	0.66	0.67
<b>NB</b>	0.75	0.77	0.75	0.75
<b>RF</b>	0.76	0.77	0.76	0.76
<b>DT</b>	0.71	0.71	0.71	0.71
<b>SVM</b>	0.76	0.77	0.76	0.77

Table VII shows that accuracy, precision, recall, and F1 measure had decreased after removing the big five personality traits from the dataset, compared to the results in Table I.

Also, Table VII shows that the SVM provides the best performance in terms of accuracy, precision, recall, and F1 measure.

On the other hand, Table I shows that the decision tree technique provides the best performance in terms of accuracy, precision, recall, and F1 measure.

This emphasizes that the existence of the big five personality traits has a very significant role in predicting a student’s academic performance.

B. The Impact of Removing the IQ Feature

Another experiment has been conducted to evaluate the performance of all techniques after removing the IQ feature from the dataset. This experiment was performed to discover the extent of the impact of the IQ feature on the academic performance of students. The results are presented in Table VIII.

Table VIII shows that the decision tree technique provides the best performance in terms of accuracy, precision, recall, and F1 measure; this corresponds to the results in Table I.

Table VIII shows that accuracy, precision, recall, and F1 measure had a little decrease after removing the IQ feature from the dataset, compared to the results in Table I.

This means that the existence of the IQ feature has an important role in the prediction of student academic performance in the proposed prediction approach.

TABLE VIII. PREDICTION RESULTS WITHOUT IQ FEATURE

Technique	Accuracy	Precision	Recall	F1 measure
KNN	0.76	0.76	0.76	0.76
NB	0.84	0.83	0.84	0.83
RF	0.82	0.81	0.82	0.82
DT	0.89	0.88	0.89	0.88
SVM	0.86	0.86	0.87	0.86

After conducting several experiments, the results have proved that “conscientiousness” is the most significant predictor when it is integrated with IQ.

To conclude, it was found that merging the big five personality traits with IQ and the academic feature besides the other moderator features namely, gender, region/city, and the number of brothers/sisters provided better results in all performance measures (accuracy, precision, recall, and F1 measure) compared to each one of them separately.

A comparison of accuracy with adding and removing the important features for all applied machine learning techniques is shown in Fig. 5.

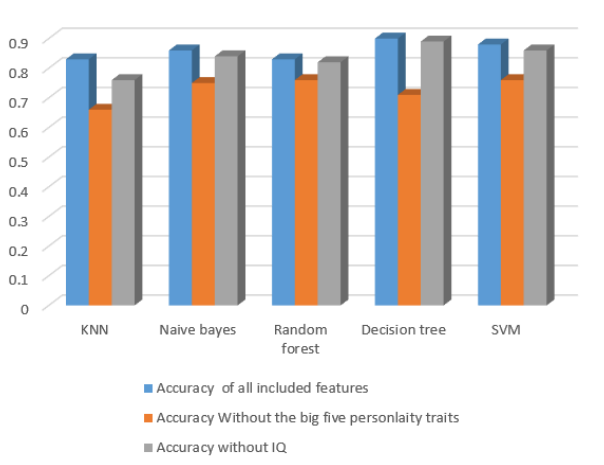


Fig. 5. Accuracy Comparison of All Techniques with and without Important Features.

### VIII. CONCLUSION AND FUTURE WORK

Predicting student academic performance is very helpful for educators and learners to improve the teaching and learning processes. In this paper, student academic performance was predicted by applying various machine learning techniques with different features.

The main idea of this research is to discover the impact of the student’s personality and the IQ along with other moderator features namely, gender, region/city, and the number of brothers/sisters integrated with the student’s GPA on the student’s academic performance.

Classification algorithms are widely used in education. K-nearest neighbor, decision trees, support vector machine,

random forests, and Naive Bayes techniques were used to predict the student’s academic performance. The decision tree technique performed the best in predicting the student’s academic performance.

To enhance the effectiveness of the proposed approach, the personality traits and the IQ features were removed. After that, the proposed approach was re-implemented after removing these features then the performance measures were re-computed. The results showed a decrease in accuracy, precision, recall, and F1 measure which emphasizes the significant role of personality traits and IQ in predicting student academic performance.

To conclude, a lot of work can be done in predicting student academic performance so further research can be conducted as future work. This helps the educational systems to track the student’s academic performance in a structured way. Furthermore, further research can use deep learning techniques and neural networks besides machine learning techniques to enhance the results.

### ACKNOWLEDGMENT

We would like to express our appreciation to the students in the Faculty of Computers and Artificial Intelligence, Cairo University for their valuable role in helping us to collect the data needed for this research work.

### REFERENCES

- [1] K. Phanniphong, P. Nuankaew, D. Teeraputon, W. Nuankaew, M. Boontongle and S. Bussaman “Clustering of learners performance based on learning outcomes for finding significant courses”, *In Proc. of ECTI DAMT-NCON - 4th International Conference on Digital Arts, Media and Technology and 2nd ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering*, Nan, Thailand, doi: 10.1109/ECTINCON. 2019. 8692263, 2019, pp. 192-196.
- [2] S. Bussaman, W. Nuankaew, P. Nuankaew, N. Rachata, K. Phanniphong and P. Jedeejit. “Prediction models of learning strategies and learning achievement for lifelong learning”, *In Proc. of IEEE International Conference on Teaching, Assessment and Learning for Engineering*, Wollongong, Australia, doi: 10.1109/TALE.2017.8252332, January 2018, pp. 192-197.
- [3] Nuankaew, Wongpanya, and J. Thongkam, “Improving student academic performance prediction models using feature selection”, *In Proc. of 17th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications, and Information Technology (ECTI-CON)*, Phuket, Thailand, 2020, pp. 392-395.
- [4] S. Hirokawa, “Key attributes for predicting student academic performance”, *In Proc. of the 10th International Conference on Education Technology and Computers*, New York, USA, DOI: <https://doi.org/10.1145/3290511.3290576>, 2018, pp. 308-313.
- [5] Katarya, Rahul, J. Gaba, A. Garg, and V. Verma. "A review on machine learning based student’s academic performance prediction systems”, *In Proc. of International Conference on Artificial Intelligence and Smart Systems (ICAIS)*, India, 2021, pp. 254-259.
- [6] A. A Aziz and N. H. I. F. Ahmad, “First semester computer science students academic performances analysis by using data mining classification algorithms”, *In Proc. of the International Conference on Artificial Intelligence and Computer Science (AICS)*, Bandung, Astanaanyar, 2014, pp. 15-16.
- [7] E. B. Costa, B. Fonseca, M. A. Santana, F. F. de Araújo, and J. Rego, “Evaluating the effectiveness of educational data mining techniques for early prediction of students’ academic failure in introductory programming courses”, *Computers in Human Behavior*, vol. 73, doi: 10.1016/j.chb.2017. pp. 247-256, 2017.



- [8] A. Dutt and M. A. Ismail, "Can we predict student learning performance from lms data? a classification approach", *In Proc. of 3rd International Conference on Current Issues in Education*, Yogyakarta, Indonesia, 2019, pp. 24-29.
- [9] Z. Cai, X. Fan, and J. Du, "Gender and attitudes toward technology use: A meta-analysis", *Computers and Education*, vol. 105, pp. 1-13, 2017.
- [10] C. J. Asarta and J. R. Schmidt, "Comparing student performance in blended and traditional courses: Does prior academic achievement matter?", *The Internet and Higher Education*, vol. 32, pp. 29-38, 2017.
- [11] M. Hussain, W. Zhu, W. Zhang, and S. M. R. Abidi, "Student engagement predictions in an e-learning system and their impact on student course assessment scores", *Computational Intelligence and Neuroscience*, vol. 2018, pp. 1-21, 2018.
- [12] E. A. Amrieh, T. Hamtini, and I. Aljarah, "Mining educational data to predict student's academic performance using ensemble methods", *International Journal of Database Theory and Application*, vol. 9, no. 8, pp. 119-136, 2016.
- [13] A. S. Hashim, W. A. Awadh, and A. K. Hamoud, "Student performance prediction model based on supervised machine learning algorithms", *In Proc. of IOP Conference Series: Materials Science and Engineering*, Thi-Qar, Iraq, 2020.
- [14] H. Waheed, S.-U. Hassan, N. R. Aljohani, J. Hardman, S. Alelyani, and R. Nawaz, "Predicting academic performance of students from big data using deep learning models", *Computers in Human Behavior*, vol. 104, p. 106-189, 2020.
- [15] K. Mongkhonvanit, K. Kanopka, and D. Lang, "Deep knowledge tracing and engagement with moocs". *In Proc. of the 9th International Conference on Learning Analytics and Knowledge*, Tempe, Arizona, 2019, pp. 340-342.
- [16] J. C. S. Silva, J. L. Ramos, R. L. Rodrigues, A. S. Gomes, F. d. F. de Souza, and A. M. A. Maciel, "An edm approach to the analysis of students' engagement in online courses from constructs of the transactional distance", *In Proc. of IEEE 16th International Conference on Advanced Learning Technologies (ICALT)*, Austin, TX, USA, 2016, pp. 230-231.
- [17] Mammadov and Sakhavat, "Big Five personality traits and academic performance: A meta-analysis", *Journal of Personality*, vol. 90, no. 2, pp.222-255, 2022.
- [18] Krishnamoorthy and Shivsubramani, "Student performance prediction, risk analysis, and feedback based on context-bound cognitive skill scores", *Education and Information Technologies* vol. 27, no. 3, pp. 3981-4005, 2022.
- [19] Buraimoh, Eluwumi, R. Ajoodha, and K. Padayachee. "Application of machine learning techniques to the prediction of student success", *In Proc. of IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS)*, Toronto, Canada, 2021, pp. 1-6.
- [20] G. Shidaganti, I. Yadav, and H. Dagdi. "Identification of Student Group Activities in Educational Institute Using Cognitive Analytics", *In Proc. of International Conference on Innovative Computing and Communications* Springer, Singapore, 2022, pp. 275-284.
- [21] Prateek, Student Group Activity Recognition, GitHub repository. <https://github.com/prateekiest/Student-Group-Activity-Recognition>, 2017.
- [22] S. Boran, K. Dimililer, and K. Tuncal, "Student performance prediction and classification using machine learning algorithms", *In Proc. of the 8th International Conference on Educational and Information Technology*, Cambridge, UK, 2019, pp. 7-11.
- [23] N. S. Altman, "An introduction to kernel and nearest-neighbor nonparametric regression", *The American Statistician*, vol. 46, no. 3, pp.175-185, 1992.
- [24] Wang, Lipo, ed, "Support vector machines: theory and applications", Springer Science and Business Media, vol. 177, 2005.
- [25] Myles, J. Anthony, N.Robert, Feudale, Y. L., Nathaniel A. Woody, and Steven D. Brown. "An introduction to decision tree modeling", *Journal of Chemometrics: A Journal of the Chemometrics Society*, vol. 18, no. 6, pp.275-285, 2004.
- [26] Biau, Gérard, and E. Scornet, "A random forest guided tour", vol. 25, no. 2, pp.197-227, 2016.
- [27] Rish and Irina, "An empirical study of the naive Bayes classifier", *In IJCAI 2001 workshop on empirical methods in artificial intelligence*, vol. 3, no. 22. pp. 41-46, 2001.
- [28] "Feature importances with a forest of trees". [Online]. Available: [https://scikit-learn.org/stable/auto\\_examples/ensemble/plot\\_forest\\_importances.html](https://scikit-learn.org/stable/auto_examples/ensemble/plot_forest_importances.html) [Accessed: 11-June-2022].
- [29] "Confusion matrix for multi-class classification". [Online]. Available: <https://www.analyticsvidhya.com/blog/2021/06/confusion-matrix-for-multi-class-classification/> [Accessed: 14-July-2022].



# Real Time Fire Detection using Color Probability Segmentation and DenseNet Model for Classifier

Faisal Dharma Adhinata<sup>1</sup>

Faculty of Informatics  
Institut Teknologi Telkom Purwokerto  
Purwokerto, Indonesia

Nur Ghaniaviyanto Ramadhan<sup>2</sup>

Faculty of Informatics  
Institut Teknologi Telkom Purwokerto  
Purwokerto, Indonesia

**Abstract**—The forest is an outdoor environment not touched by the surrounding community, so it is not immediately handled when a fire occurs. Therefore, surveillance using cameras is needed to see the presence of fire hotspots in the forest. This study aims to detect hotspots through video data. As is known, fire has a variety of colors, ranging from yellow to reddish. The segmentation process requires a method that can recognize various fire colors to get a candidate fire object area in the video frame. The methods used for the color segmentation process are Gaussian Mixture Model (GMM) and Expectation–maximization (EM). The segmentation results are candidates for fire areas, which in the experiment used the value of  $K=4$ . This fire object candidate needs to be ascertained whether the segmented object is a fire object or another object. In the feature extraction stage, this research uses the DenseNet-169 or DenseNet-201 models. In this study, various color tests were carried out, namely RGB, HSV, and YCbCr. The test results show that RGB color produces the most optimal training accuracy. This RGB color configuration is used to test using video data. The test results show that the true positive and false negative values are quite good, 98.69% and 1.305%. This video data processing produces fps with an average of 14.43. So, it can be said that this combination of methods can be used to process real time data in case studies of fire detection.

**Keywords**—Fire detection; color segmentation; GMM-EM; DenseNet; real time

## I. INTRODUCTION

Human daily life cannot be separated from the heat energy produced by a fire. The heat energy from this fire is often used for cooking, lighting candles as a light source, and burning garbage. However, fires can be catastrophic if they are not controlled and burn a large area. Fires can occur in indoor and outdoor environments such as forests. In Indonesia, forest fires often occur in Sumatra and Kalimantan because forest areas are still common [1]. Natural factors and human error can cause the emergence of fire hotspots. Some natural factors are hot weather, wind, and chemical reactions [2]. Then human error can occur due to forgetfulness in activities with fire, especially in rural areas that still use firewood for daily life [3]. Currently, the government has made efforts to mitigate fire disaster management [4], but the efforts made have not used Artificial Intelligence technology for automation. Therefore, the need for prevention efforts by detecting hotspots as early as possible before the fire spreads. This hotspot detection process can be done by installing an intelligent camera programmed using Artificial Intelligence to identify hotspots.

Several researchers have developed early detection of hotspots, including fire detection using video [5] and sensors [6]. Limitations in using sensors, especially gas sensors, can occur when there is other smoke, for example, people smoking. Then the heat sensor can also go wrong when the weather is hot. Using fire detectors through sensors also costs a lot when used in an outdoor environment because they must replicate the tool at many points. So, the proposed research focuses on using video to detect hotspots. Video data can be obtained by installing a Closed-Circuit Television (CCTV) camera. CCTV camera can detect fires using digital image processing and computer vision technologies, known as image-based fire detection. The advantages of image-based fire detection compared to conventional fire detectors can be installed in a large, open area to reduce expenses. The use of video data requires a method that can run in real-time [7]. Besides, the video resolution also affects the detection accuracy results. In a previous study [2], the fire detection system produced a reasonably accurate accuracy, but the processing time of each frame could not be done in real time. It is also a limitation of previous research. The main steps that affect the speed and accuracy are image segmentation, feature extraction and classification.

The segmentation process is carried out to take the fire area in the video frame. The segmentation stage of searching candidate fire object is very important to separate the fire candidate object from the background, which should not enter the feature extraction stage. The color of fire is a combination of various colors, ranging from reddish to yellow [8]. Previous studies conducted experiments using fire color segmentation, including RGB, HSV, and YCbCr color features, have not produced optimal accuracy [9]. The lack of this feature is because the color of the fire changes due to the wind. Therefore, this research uses a segmentation method that can overcome the quick color change of the fire using probability. This probability makes several color combinations of fire. The proposed research uses color probabilities to perform segmentation. In other case study research, the segmentation process was carried out on the image using a combination of the Gaussian Mixture Model (GMM) and Expectation–maximization (EM) methods [10][11]. The segmentation results show that combining these methods can detect multi-colored objects. Therefore, the proposed research uses of GMM-EM for the segmentation of candidate fire objects contained in video frames.

After the candidate fire object is obtained, the fire object must be sure that the segmented one is a fire object. Several studies of feature extraction and classification use the transfer learning method. The transfer learning of DenseNet201 model is used for image classification [12]. The results showed good accuracy for the feature extraction of corn disease. In another research, the DenseNet model was also used for feature extraction of the lungs affected by Covid-19 [13]. The results showed good accuracy using the DenseNet model. This research will also use the DenseNet model at the feature extraction and classification stage. The result of this research is a real-time fire early detection system using video data.

This research aims to build a real-time fire point detection system using video data for early warning of fires. Speed is an important thing in this study to be evaluated. The color of fire that is not only yellow requires a precise segmentation process, so the proposed method uses a combination of various colors of fire. The segmentation results are then extracted and classified to ensure that the object is a fire. Overall, the contributions of this research are:

- The use of various color combinations of fire to perform the segmentation process for searching fire object candidates.
- Evaluation of the segmentation process on each video frame to minimize non-fire object detection errors.
- The use of transfer learning as feature extraction and classification to achieve optimal accuracy and real-time processing.

## II. RELATED WORK

The development of fire detection applications often uses sensors [14]. The downside of using this sensor depends on the surrounding weather. When using a heat sensor during the dry season, the sensor may experience error detection. Then the gas sensor can also experience an error when there is other smoke, for example, cigarette smoke, smoke from burning garbage, etc. Even detecting hotspots in open areas, such as forests, is very difficult. Therefore, the fire detection uses video data.

Several researchers who process fire video data, including Khan et al. [15], used a fire's color, perimeter, area, and roundness for an indoor fire case study. The method used does not consider small fires, so it cannot carry out early detection of hotspots. Then, research by Thepade et al. [9] used a color combination of HSV and YCbCr to detect hotspots. The method used is still static, so the use of dynamic video data cannot be done. The segmentation process can also use the deep feature [16]. This deep feature is suitable for high-resolution images such as satellite images. Several segmented objects produce relatively good accuracy. However, the disadvantage of using deep features is that the processing time is quite long, so it is unsuitable for real-time processing. The color component of fire is not only red, but a combination of various colors, including yellow, orange, red, white, and blue. Previous research by Dong Keun Kim [10] used Gaussian Mixture Model (GMM) and Expectation-maximization (EM) to detect color combinations on objects. The proposed research will segment fire objects with fire color data training using the

GMM-EM method. Video resolution also affects the detection accuracy results. In a previous study [2], the fire detection system produced a fairly accurate accuracy of 99.7%, but the processing time of each frame took 0.23 seconds or four fps. Therefore, this research proposes a new approach to obtain optimal accuracy and can run in real-time.

Currently, deep learning is a method that researchers often use for classification case studies. Deep learning, frequently used to handle picture data, is called Convolutional Neural Network. Deep learning is a technique used by artificial neural networks to manage input data utilizing multiple hidden layers. The output of this process is a non-linear modification of the input data used to determine the output value [17]. Deep learning is typically used for vast amounts of data. However, the data is relatively small in some instances, such as in this fire detection scenario. Transfer learning is a strategy for processing small amounts of data in which the model has been trained using other data [18]. DenseNet is an example of a transfer learning model. In this research, an evaluation of the DenseNet-169 and DenseNet201 models will be carried out.

## III. PROPOSED METHOD

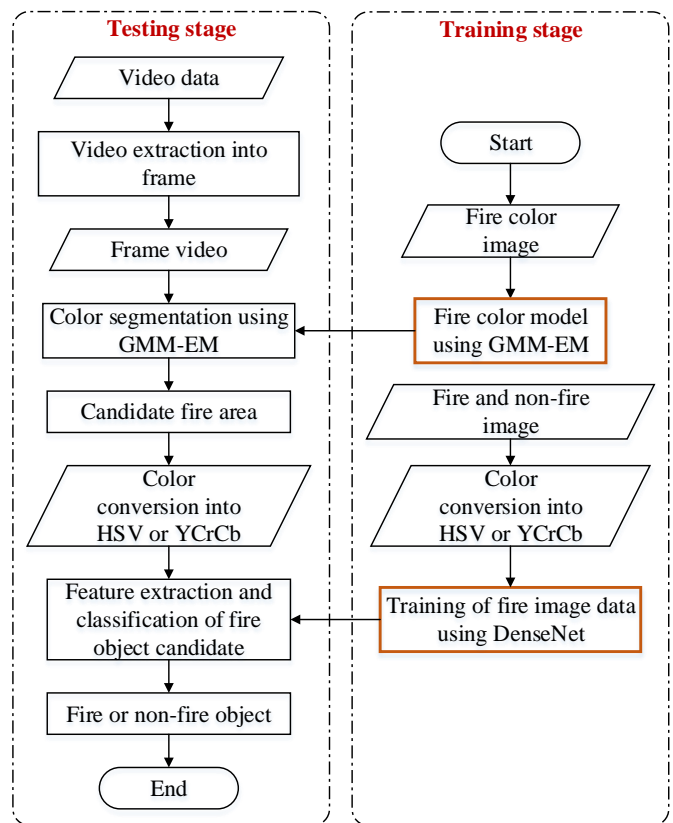


Fig. 1. Proposed System of Fire Detection.

The hotspot detection system starts from the training stage. There are two training processes: training for the segmentation process and feature extraction on fire objects. The training process uses a combination of GMM and EM methods. The fire segmentation process uses image data of fire colors. Then, the feature extraction process uses fire and non-fire image data. Before the training process, the image is converted to HSV or YCbCr color. The training process uses the transfer learning

method. The best model is used for matching video data. Then the testing phase begins with real-time video data input. Video data is extracted in the form of frames. Detection of fire candidates in video frames is done by matching the color model. Flame object candidates are converted to HSV or YCbCr color. The conversion results are matched with the feature extraction model. The feature extraction and classification stage use DenseNet model. The results of the classification are fire and non-fire objects. Fig. 1 shows the flowchart of this research's fire point detection.

### A. Acquisition Data

This research used two datasets: a dataset for segmentation of fire object candidates and a dataset for fire object classification. The dataset for segmentation uses 30 images of fire color images. This dataset uses three color channels: Red, Green, and Blue (RGB), measuring 100 x 100. The features of this dataset were taken from it based on the RGB color model, which was used to show the different colors of fire in the color probability model. The fire color varies so that it can detect various colors of fire when testing using video data. Fig. 2 shows an example of a fire color dataset used for the segmentation process.



Fig. 2. Fire Color Data for Segmentation Stage.

Then at the feature extraction stage, the data uses from Kaggle created by Jadon et al. [19]. This dataset consists of two classes, namely the fire class in various places and the non-fire class like other objects. The number of fire data is 1123 images, while the non-fire data is 1301. It was made by taking images of fire and things that don't fire under challenging situations, like the fire image in the forest and the non-fire image with things that look like fire in the background. At the training stage, the percentage of training data used is 80%, while the testing data is 20%. Fig. 3 shows an example of training data for the feature extraction stage.

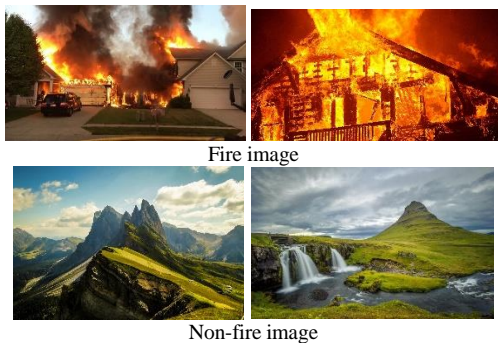


Fig. 3. Fire and Non-Fire Image for Feature Extraction Stage.

### B. Fire Object Segmentation

Multiple clusters can describe a dataset's distribution. Modelling a dataset with a single mean (one Gaussian) and estimated parameters is not optimal. For example, if a dataset contains two means of 218 and 250, the average may be close

to 221. It is not a precise estimate. Multiple Gaussians with means of 218 and 250 provide a more accurate representation of the distribution of the data set.

In situations where multiple data sets with varying numbers of clusters describe the same feature, it is preferable to model the data across the three sets using a multivariate Gaussian [20]. Equation (1) represents the multivariate Gaussian equation. It allows for a more precise evaluation of the distribution of clusters across the provided data.

$$N(x|\mu, \Sigma) = \frac{1}{(2\pi|\Sigma|)^{\frac{1}{2}}} \exp \left\{ -\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right\} \quad (1)$$

This research employs a multivariate Gaussian with three color channels: Red, Green, and Blue. It has detected fire-based objects, so the number of clusters in each color channel will be examined. This research uses the Expectation-Maximization algorithm to estimate the means and covariances and determine the probability of a pixel belonging to a cluster. The total image is then modelled with a three-dimensional Gaussian.

The following sections outline the stages involved in performing the EM algorithm:

1) *Using some random numbers:* initialize the means and covariances. The covariance matrix must have the shape (dim, dim), where dim is the Gaussian's dimension number. These values are stored in a dictionary data structure called 'parameters.'

2) *E Stage:* Gaussians are combined in (2).  

$$p(x) = \sum_{k=1}^K \pi_k N(x|\mu_k, \Sigma_k) \quad (2)$$

These are the probabilities associated with a given value x. It can accomplish this by applying the Bayes rule as (3).

$$= \frac{\pi_k N(x|\mu_k, \Sigma_k)}{\sum_{j=1}^K \pi_j N(x|\mu_j, \Sigma_j)} \quad \text{where, } \pi_k = \frac{N_k}{N} \quad (3)$$

These are saved in an array named 'cluster prob' with the dimensions (n\_feat, K). n\_feat is the number of rows in the dataset in this case.

3) *M Stage:* Then, update the means and covariances. This can be accomplished using the following (4).

$$\mu_j = \frac{\sum_{n=1}^N \gamma_j(x_n) x_n}{\sum_{n=1}^N \gamma_j(x_n)}$$

$$\Sigma_j = \frac{\sum_{n=1}^N \gamma_j(x_n) (x_n - \mu_j)(x_n - \mu_j)^T}{\sum_{n=1}^N \gamma_j(x_n)}$$

$$\pi_j = \frac{1}{N} \sum_{n=1}^N \gamma_j(x_n) \quad (4)$$

4) *Calculate the Log Likelihood:* The objective is to increase the log likelihood function until the change in likelihood is equal to or less than a specified number. The following (5) is used to get the log likelihood.

$$\ln p(X|\mu, \Sigma, \pi) = \sum_{n=1}^N \ln \left\{ \sum_{k=1}^K \pi_k N(x_n|\mu_k, \Sigma_k) \right\} \quad (5)$$

The log likelihood is appended to the log likelihoods array. The log probability difference is calculated by subtracting the most recent value from the most recently stored value. This technique is continued iteratively until the difference in log likelihood reaches a predefined value or the maximum number

of iterations is reached. The final values produced are the dataset's estimated means and covariances. Fig. 4 shows an example of segmentation results.

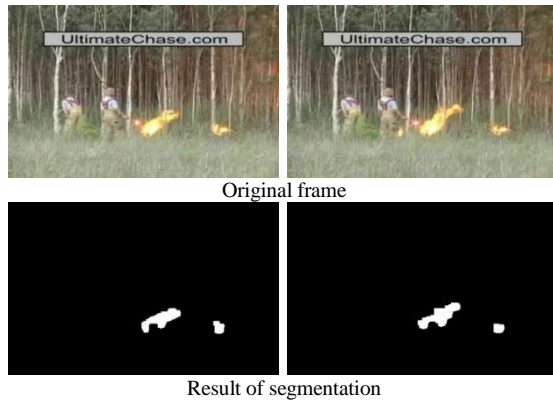


Fig. 4. Fire and Non-Fire Image for Feature Extraction Stage.

### C. Feature Extraction and Classification on Fire Object Candidate

In this study, the feature extraction phase uses HSV or YCbCr colors. The image of the flame candidate is converted to HSV or YCbCr color space. The results of this conversion are utilized in the process of feature extraction. The process of feature extraction employs the DenseNet transfer learning model [21]. A DenseNet is a convolutional neural network with dense connections between layers using Dense Blocks, where all layers (with matching feature-map sizes) are directly connected. Maintain the feed-forward nature; each layer receives additional inputs from all preceding levels and sends its feature maps to all subsequent levels.

Model: "sequential"

Layer (type)	Output Shape	Param #
densenet201 (Functional)	(None, 1, 1, 1920)	18321984
flatten (Flatten)	(None, 1920)	0
dense (Dense)	(None, 256)	491776
dense_1 (Dense)	(None, 1)	257
-----		
Total params:	18,814,017	
Trainable params:	492,033	
Non-trainable params:	18,321,984	

Fig. 5. Architecture for Training Stage on Fire and Non-Fire Image.

At the training stage, the preprocessing process such as resizing to 224x224 pixels and normalizing the data after resizing the data. The training data image uses a method with the DenseNet model, which includes feature extraction and classification processes. The DenseNet model has general operations for batch normalization, ReLU activation, and convolution. DenseNet model with 201 layers has dense block 1, transition layer 1, dense block 2, transition layer 2, dense block 3, transition layer 3, dense block four, and classification layer processes that produce the output model with .h5 format. In this research, feature extraction will be carried out using the

DenseNet-169 and DenseNet-201 models. Fig. 5 shows the architectural configuration used for the training process.

### D. System Evaluation

Using video data in a fire detection system requires evaluation, especially regarding the accuracy and speed of processing video frames. The first test is carried out at the segmentation stage. The segmentation process is used to find the fire object candidate area in the video frame. This research evaluates the value of  $K$  used in the GMM method against the segmentation results and the resulting fps. Then at the feature extraction stage, this research assesses the use of RGB, HSV, and YCbCr colors to see the results of training accuracy. The last configuration is used for evaluation using video data to see the true positive and false negative values of the video data matching results. The last is an evaluation of the video data processing speed to see the fps value.

## IV. RESULT AND DISCUSSION

This section conducts some experiments at the segmentation stage, feature extraction and classification of fire or non-fire objects. The last is matching using video data. In this experiment, this research used a computer with Core i5 specifications with 8GB of RAM and VGA GTX 1650. Then, the computer program uses Python programming language.

### A. Fire Object Segmentation

At the segmentation stage using the GMM and EM methods, the most influential parameter is  $K$  value, which functions as a clustering dataset of fire colors. Table I shows the results of the variation of the  $K$  value on the segmentation results.

TABLE I. THE EFFECT OF K VALUE ON SEGMENTATION RESULTS

K Value of GMM	Ground Truth Image	Segmentation on Frame	Fps
2			31.28
3			29.22
4			20.66
5			17.65
6			15.13

Based on the experiment using the  $K$  value in the GMM method, there are no segmented fire objects when the value of  $K = 2$ . Whereas in the ground truth image, there are two fire objects contained in the image. Then at the value of  $K = 3$  to

6, the segmentation results show two fire objects with the same ground truth. However, if it looks closely, the more  $K$  values are added, the closer the segmentation results get to the ground truth shape. In processing video data also need to pay attention to the resulting speed. In this experiment, it was tested with the resulting fps value. As the value of  $K$  increases, the resulting fps also decreases. Because the number of clusters is increasing, it takes time to match each cluster. Therefore, this research choses a value of  $K = 4$ , which still produces an average of 20 fps. This configuration will be used to test using multiple videos containing fire objects.

### B. Feature Extraction and Classification on Candidate Fire Object

This transfer learning model uses to process features into the feature extraction layer before the classification layer. The feature extraction used in this study is the DenseNet-169 or DenseNet201 model. The difference between DenseNet-169 and DenseNet-201 is the number of parameters. In DenseNet-169 it is 14.3M, while in DenseNet-201, it is 20.2M [22]. This study's training process configuration uses an image input size of 50 x 50. Then the distribution of training data and testing data is 80% training data and 20% testing data. Then the optimizer used is Adam with a loss configuration using binary\_crossentropy because the number of classes used is two, namely fire and non-fire. Table II shows training results using two DenseNet models by monitoring validation accuracy. This research experimented with three colors, namely RGB, HSV, and YCbCr.

TABLE II. THE TRAINING RESULT FOR FEATURE EXTRACTION STAGE

Epoch	Transfer Learning Model					
	DenseNet-169			DenseNet-201		
	RGB	HSV	YCbCr	RGB	HSV	YCbCr
1	0.8438	0.6484	0.6562	0.8672	0.6484	0.6406
2	0.9766	0.8750	0.8906	0.9531	0.8984	0.8047
3	0.9844	0.9453	0.9531	0.9766	0.8984	0.8750
4	0.9844	0.9531	0.9375	0.9844	0.9375	0.9219
5	1.0000	0.9688	0.9609	0.9922	0.9375	0.9453
6	1.0000	0.9844	0.9609	1.0000	0.9688	0.9531
7	1.0000	-	0.9766	1.0000	0.9844	0.9766
8	1.0000	-	0.9844	1.0000	0.9922	0.9766

Based on the experimental results in Table II, the best results are obtained using RGB colors. It is because the pre-trained model uses images with RGB colors in the transfer learning model. So, when tested using other colors such as HSV and YCbCr, the accuracy results obtained have not reached 100% in epoch 8. This training process uses an early stop with a maximum of no change of 5 epochs. In this experiment, all models stopped at the eighth epoch. Therefore, this research used RGB color as the color configuration in the video data experiment. From the DenseNet-169 and DenseNet-201 models, the best results are obtained using the DenseNet-169 model because in the fourth epoch, the accuracy is 100%, and the DenseNet-169 model is lighter, which affects faster data processing. Therefore, in the experiment using video data, this research used the DenseNet-169 model.

### C. Matching with Video Data

The segmentation and feature extraction models were obtained for video data testing. In this test, the data used is a

fire video obtained from the VisiFire fire detection software [23]. All video datasets have a resolution of 400 x 256 at 15 fps. The number of video frames varies, Controlled1 260 frames video, Controlled2 246 frames, Controlled3 208 frames, Forest1 200 frames, Forest2 245 frames, and Forest3 255 frames. It will check whether a fire object is detected in the video frame. It will evaluate true positive (TP), and false negative (FN) results in each video experiment. Table III shows the results of the evaluation of video data processing.

TABLE III. RESULT OF MATCHING WITH VIDEO DATA

Video	Proposed Method		Color + SVM [24]		Tempo-spatial + SVM [25]	
	TP	FN	TP	FN	TP	FN
Controlled1	100	0	55.2	44.8	94.98	5.02
Controlled2	100	0	77.7	22.3	-	-
Controlled3	100	0	97.9	2.1	95	5
Forest1	100	0	-	-	-	-
Forest2	100	0	-	-	-	-
Forest3	92.17	7.83	-	-	-	-
<b>Average</b>	<b>98.69</b>	<b>1.305</b>	<b>76.93</b>	<b>23.067</b>	<b>94.99</b>	<b>5.01</b>

The experimental results show that the combination of segmentation and feature extraction models produces a reasonably good true positive, 98.69%. In previous studies, 95% true positive results did not exist in the model using supervised learning. Likewise, with false negative results, in this study, the value was below 2 percent, which means that only a few fire objects were not detected. Previous research also used the handcrafted method, which means that the features obtained are based on the components contained in the fire object. The classification process is also carried out using machine learning. Quantitatively, the average true positive of the proposed method is better than the previous research. The amount of video data tested is also more, so this method passes more test data with various fire object conditions. In this case, it makes qualitative testing of the proposed method better. In addition to testing true positive and false negative values, we also evaluate the resulting fps results for video data processing. Table IV shows the fps results obtained from the tested videos.

TABLE IV. RESULT OF COMPUTATION TIME

Video	Fps
Controlled1	16.63
Controlled2	14.75
Controlled3	11.42
Forest1	14.84
Forest2	12.6
Forest3	16.36
<b>Average</b>	<b>14.43</b>

The video used to test the fps is 400x 256 resolutions. The fps results obtained based on Table IV are not the same because the fire objects detected in the video are different. The more fire objects there are in a frame, the fps result also decreases. The average fps produced is quite good, namely 14.43 fps, meaning that for 1 second, it can process around 14 frames. An example of the results of the segmentation and detection processes in this system is shown in Fig. 6. This study has limitations related to the resulting fps that are not optimal. There are still about 12 fps in testing, while CCTV cameras usually produce 15 fps recordings. The challenge in further research is to increase the resulting fps value so that the



use of CCTV cameras with high fps can be applied. However, video data processing must also consider the detection results in addition to the resulting fps value. In this study, the number of true positives produced was quite good, 98.69%.

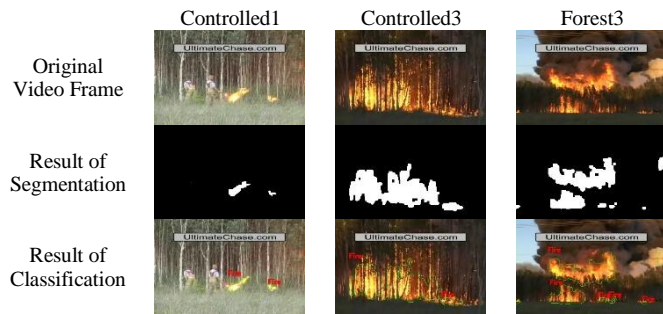


Fig. 6. Example Fire Detection on Video Frame.

## V. CONCLUSION

Fire is a disaster that must be handled immediately so that it does not spread to a broader area. Early hotspot detection is needed, so a fire is directly identified to extinguish the fire. This research proposes a framework for fire detection using video data. The detection process starts with the fire object candidate segmentation. The fire object candidate area was performed by feature extraction and classification using the DenseNet model. It matches results using video data, resulting in true positive values of 98.69% and 14.43 fps. Future research can modify the combination of segmentation and feature extraction methods to produce higher fps. It is because with the development of technology, of course, CCTV cameras will also produce clearer videos with more fps. Therefore, future research is still very open to improving the resulting fps for real-time processing. In addition, if the method produces a high enough fps, it can be applied for implementation with configurations through embedded system devices with CCTV cameras.

## ACKNOWLEDGMENT

We would like to thank the Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi Indonesia through the Penelitian Dosen Pemula (PDP) grant at the number of IT Tel3701/LPPM-000/Ka.LPPM/VI/2022.

## REFERENCES

- [1] R. B. Edwards, R. L. Naylor, M. M. Higgins, and W. P. Falcon, "Causes of Indonesia's forest fires," *World Dev.*, vol. 127, 2020, doi: 10.1016/j.worlddev.2019.104717.
- [2] A. Abdusalomov, N. Baratov, A. Kutlimuratov, and T. K. Whangbo, "An improvement of the fire detection and classification method using YOLOv3 for surveillance systems," *Sensors*, vol. 21, no. 19, 2021, doi: 10.3390/s21196519.
- [3] R. L. Naylor, M. M. Higgins, R. B. Edwards, and W. P. Falcon, "Decentralization and the environment: Assessing smallholder oil palm development in Indonesia," *Ambio*, vol. 48, no. 10, pp. 1195–1208, 2019, doi: 10.1007/s13280-018-1135-7.
- [4] A. A. Fitriany, P. J. Flatau, K. Khoirunurrofik, and N. F. Riama, "Assessment on the use of meteorological and social media information for forest fire detection and prediction in riau, indonesia," *Sustain.*, vol. 13, no. 20, 2021, doi: 10.3390/su13201188.
- [5] F. Gong et al., "A real-time fire detection method from video with multifeature fusion," *Comput. Intell. Neurosci.*, vol. 2019, 2019, doi: 10.1155/2019/1939171.

- [6] R. S. Kharisma and A. Setiyansah, "Fire early warning system using fire sensors, microcontroller, and SMS gateway," *J. Robot. Control*, vol. 2, no. 3, pp. 165–169, 2021, doi: 10.18196/jrc.2372.
- [7] A. Nurhopipah and A. Harjoko, "Motion Detection and Face Recognition for CCTV Surveillance System," *IJCCS (Indonesian J. Comput. Cybern. Syst.*, vol. 12, no. 2, p. 107, 2018, doi: 10.22146/ijccs.18198.
- [8] R. Sadek et al., "Novel colored flames via chromaticity of essential colors," *Def. Technol.*, vol. 15, no. 2, pp. 210–215, 2019, doi: 10.1016/j.dt.2018.05.002.
- [9] S. D. Thepade, J. H. Dewan, D. Pritam, and R. Chaturvedi, "Fire Detection System Using Color and Flickering Behaviour of Fire with Kekre's LUV Color Space," *Proc. - 2018 4th Int. Conf. Comput. Commun. Control Autom. ICCUBEA 2018*, pp. 1–6, 2018, doi: 10.1109/ICCUBEA.2018.8697454.
- [10] D. K. Kim, "Color detection using Gaussian mixture model," *J. Theor. Appl. Inf. Technol.*, vol. 95, no. 17, pp. 4313–4320, 2017.
- [11] Y. Li, J. Zhang, R. He, L. Tian, and H. Wei, "Hybrid DE-EM Algorithm for Gaussian Mixture Model-Based Wireless Channel Multipath Clustering," *Int. J. Antennas Propag.*, vol. 2019, pp. 1–10, 2019, doi: 10.1155/2019/4639612.
- [12] F. D. Adhinata, G. F. Fitriana, A. Wijayanto, M. Pajar, and K. Putra, "Corn Disease Classification using Transfer Learning and Convolutional Neural Network," vol. 9, no. 2, pp. 1–7, 2021.
- [13] M. K. Bohmrah and H. Kaur, "Classification of Covid-19 patients using efficient fine-tuned deep learning DenseNet model," *Glob. Transitions Proc.*, vol. 2, no. 2, pp. 476–483, 2021, doi: 10.1016/j.gltp.2021.08.003.
- [14] F. Khan, Z. Xu, J. Sun, F. M. Khan, A. Ahmed, and Y. Zhao, "Recent Advances in Sensors for Fire Detection," *Sensors*, vol. 22, no. 9, 2022, doi: 10.3390/s22093310.
- [15] R. A. Khan, J. Uddin, S. Corraya, and J.-M. Kim, "Machine vision-based indoor fire detection using static and dynamic features," *Int. J. Control Autom.*, vol. 11, no. 6, 2018, doi: 10.14257/ijca.2018.11.6.09.
- [16] S. D. Khan, L. Alarabi, and S. Basalamah, "Deep hybrid network for land cover semantic segmentation in high-spatial resolution satellite images," *Inf.*, vol. 12, no. 6, pp. 1–16, 2021, doi: 10.3390/info12060230.
- [17] A. Anton, N. F. Nissa, A. Janiati, N. Cahya, and P. Astuti, "Application of Deep Learning Using Convolutional Neural Network (CNN) Method For Women's Skin Classification," *Sci. J. Informatics*, vol. 8, no. 1, pp. 144–153, 2021, doi: 10.15294/sji.v8i1.26888.
- [18] Y. Gultom, A. M. Arymurthy, and R. J. Masikome, "Batik Classification using Deep Convolutional Network Transfer Learning," *J. Ilmu Komput. dan Inf.*, vol. 11, no. 2, p. 59, 2018, doi: 10.21609/jiki.v11i2.507.
- [19] A. Jadon, M. Omama, A. Varshney, M. S. Ansari, and R. Sharma, "FireNet: A Specialized Lightweight Fire & Smoke Detection Model for Real-Time IoT Applications," 2019, [Online]. Available: <http://arxiv.org/abs/1905.11922>
- [20] A. Ruseckaite, D. Fok, and P. Goos, "Flexible Mixture-Amount Models Using Multivariate Gaussian Processes," *J. Bus. Econ. Stat.*, vol. 38, no. 2, pp. 257–271, Apr. 2020, doi: 10.1080/07350015.2018.1497506.
- [21] T. Zhou, X. Ye, H. Lu, X. Zheng, S. Qiu, and Y. Liu, "Dense Convolutional Network and Its Application in Medical Image Analysis," *Biomed Res. Int.*, vol. 2022, p. 2384830, 2022, doi: 10.1155/2022/2384830.
- [22] I. Kousis, I. Perikos, I. Hatzilygeroudis, and M. Virvou, "Deep Learning Methods for Accurate Skin Cancer Recognition and Mobile Application," *Electron.*, vol. 11, no. 9, pp. 1–19, 2022, doi: 10.3390/electronics11091294.
- [23] A. E. Cetin, "Computer Vision Based Fire Detection Software," *VisiFire*, 2014. <http://signal.ee.bilkent.edu.tr/VisiFire>.
- [24] B. C. Ko, K.-H. Cheong, and J.-Y. Nam, "Fire detection based on vision sensor and support vector machines," *Fire Saf. J.*, vol. 44, no. 3, pp. 322–329, 2009, doi: <https://doi.org/10.1016/j.firesaf.2008.07.006>.
- [25] T. Xuan Truong and J.-M. Kim, "Fire flame detection in video sequences using multi-stage pattern recognition techniques," *Eng. Appl. Artif. Intell.*, vol. 25, no. 7, pp. 1365–1372, 2012, doi: <https://doi.org/10.1016/j.engappai.2012.05.007>.



# Tissue and Tumor Epithelium Classification using Fine-tuned Deep CNN Models

Anju T E<sup>1</sup>

Research Scholar, Dept. of Computer Science  
Mother Teresa Women's University  
Kodaikanal, India

Dr. S. Vimala<sup>2</sup>

Associate Professor, Dept. of Computer Science  
Mother Teresa Women's University  
Kodaikanal, India

**Abstract**—The field of Digital Pathology (DP) has become more interested in automated tissue phenotyping in recent years. Tissue phenotyping may be used to identify colorectal cancer (CRC) and distinguish various cancer types. The information needed to construct automated tissue phenotyping systems has been made available by the introduction of Whole Slide Images (WSIs). One of the typical pathological diagnosis duties for pathologists is the histopathological categorization of epithelial tumors. Artificial intelligence (AI) based computational pathology approaches would be extremely helpful in reducing the pathologists ever-increasing workloads, particularly in areas where access to pathological diagnosis services is limited. Investigating several deep learning models for categorizing the images of tumor epithelium from histology is the initial goal. The varying accuracy ratings that were achieved for the deep learning models on the same database demonstrated that additional elements like pre-processing, data augmentation, and transfer learning techniques might affect the models' capacity to attain better accuracy. The second goal of this publication is to reduce the time taken to classify the tissue and tumor Epithelium. The final goal is to examine and fine-tune the most recent models that have received little to no attention in earlier research. These models were checked by the histology Kather CRC image database's nine classifications (CRC-VAL-HE-7K, NCT-CRC-HE-100K). To identify and recommend the most cutting-edge models for each categorization, these models were contrasted with those from earlier research. The performance and the achievements of the proposed preprocessing workflow and fine-tuned Deep CNN models (Alexnet, GoogLeNet and Inceptionv3) are greater compared to the prevalent methods.

**Keywords**—Colorectal cancer; deep learning; CNN; tumor epithelium; Alexnet; GoogLeNet; Inceptionv3

## I. INTRODUCTION

Historically, pathologists have examined the micro-anatomy of cells and tissues under a microscope. The development of Digital Pathology (DP) imaging in recent years has given pathologists an alternative method to perform the same analysis on a computer screen [1]. The current inquiry methodologies for breast cancer include mammography, magnetic resonance imaging (MRI), and pathology examinations. The histopathological scans are recognized as a golden standard to improve the diagnostic accuracy for patients who also had other investigations, such as mammography [2]. Additionally, a histopathological examination can offer more thorough and trustworthy information to detect cancer and to evaluate, how it affects the tissues around it [3]–[5]. The new

modality, digital pathology imaging, now makes WSI (Whole Slide Imaging) a reality. Through WSI, the images may be shared, viewed on a digital display, and can be controlled/ examined on a screen [6]. Tumor architecture in Colorectal Cancer (CRC) evolves as the disease progresses [7] and is associated with patient prognosis [8]. Therefore, it is important for histopathologists to quantify the tissue composition in CRC. Inter-tumor heterogeneity and intra-tumor heterogeneity are both forms of tumor heterogeneity. By the different signals that cells pick up from their microenvironment, the tumor microenvironment (TME) really plays a significant role in the establishment of intra-tumor heterogeneity (ITH) [9]. The third most common cancer type to cause mortality is colorectal cancer (CRC), which is ranked as the fourth most common cancer [10]. In fact, treating patients and saving their lives depends on early-stage CRC diagnosis [11]. For the classification and prognostication of cancer, the study of tumor heterogeneity is crucial [12]. In-tumor heterogeneity can help to clarify, how TME affects patient prognosis and can also be used to spot new aggressive phenotypes that may be potential targets for future therapies [13]. Although most present histological analysis relies on the pathologists' subjective assessments, a critical need for automating the various processing techniques arises, that can provide good quantitative analysis and throughput of the digital pathology images for precise identification and assessment of various tumor epitheliums.

Deep convolutional neural networks (CNNs) algorithms automatically analyse images for handling classification and detection tasks, reducing the amount of manual labour necessary for the feature-extraction operations [14]. The lack of a suitably sizable annotated data set for training is a significant barrier to applying deep learning to many biological domains. Transfer learning, which makes use of deep CNNs that have already been trained on a significant amount of natural scene data, may be used to circumvent the need for sample size, nevertheless. This approach is based on the notion that the characteristics discovered by deep CNNs to identify classes in a dataset may also be useful for clinical data sets with marginally worse performance.

In medical domain there are currently three approaches in deep learning: (i). Acquiring features learned in the training phase of deep CNN with numerous natural images, then the features acquired are used for classifier training [15], [16], and [17], (ii) fine-tuning a small number of network layers are fine tuned in the pre-trained CNN on a desired data set [18], (iii)

training directly the deep CNN with real-world data. The author in [19] suggests categorizing brain tumor based on multiphase MRI scans and compares the outcomes to several deep learning structure configurations and baseline neural networks.

Based on the findings of this work, the effectiveness of identifying tumor epithelial tissues using transfer learning approaches in the area: In what ways, would fine-tuning the models in the Validation Frequency, Dropout Layer, and Classification Layer improve the classification performance, were investigated.

## II. RELATED WORKS

The classification of the different tissue types in histological images is frequently done under supervision [20]. Modern approaches for phenotyping CRC tissues under supervision can be divided into two groups: learnt methods [23,24] and methods based on texture [21]. Additionally, other efforts, like [24], integrated shallow and deep characteristics. In order to extract certain structures from image areas, hand-crafted techniques known as "texture approaches" were developed [25]. Deep learning techniques, on the other hand, have the capacity to directly learn more pertinent and sophisticated image characteristics across layers, particularly, when the relationship between both the source data as well as the expected outcomes is not known in advance. As pathological imaging activities are incredibly complex and there is little knowledge on which quantitative image properties predict the outcomes, deep learning approaches are suitable for these activities [26, 27]. In [20], Kather et al. did the primary researches to handle CRC multi-class tissue types where 5000 histological pictures were used to create a database that included eight different CRC types of tissues. Modern texture descriptors and classifiers were put to the test by J. N. Kather and colleagues. Their suggested strategy is based on a promising mix of global lower-order texture metrics along with local descriptors from GLCM and LBP. The General Purpose (GenP) approach, which Nanni et al. proposed in [24], is based on the collection of learned features, hand-crafted, and dense sampling. In their proposed method, all features were trained using SVM, and the integration were achieved using the sum rule. To differentiate between the various CRC tissue types, [28] evaluated shallow and deep characteristics. In their research, they looked at how dimensionality reduction techniques affected accuracy and computing expense. Their findings demonstrated that CNN-based features may achieve the best accuracy/dimensionality trade-off. In [29], J. N. Kather et al. created a dataset consisting of one lakh images which categorized eight tissue types using eighty-six H&E slides of CRC tissues. They evaluated the AlexNet [31,38], ResNet-50 [34], GoogLeNet [33], VGG19 [30], and SqueezeNet version-1.1 [32] pretrained CNN models. They came to the conclusion that among the five CNN models, VGG19 was the best. A novel CRC-TP database with 280K patches taken from 20 WSIs of CRC and divided into seven different tissue phenotype was proposed by Javed et al. [22]. They employed 27 cutting-edge techniques, including texture, CNN, and Graph CNN-based approaches (GCN), to categorize different tissue types. According to their test findings, the GCN performed

better than the texturing and CNN approaches. Although hand-crafted feature-based and deep learning approaches have been employed to classify many CRC tissue types, their performance still needs to be enhanced. In order to do this, deep CNN methods have been enhanced that significantly outperformed baseline results on two well-known databases.

## III. MATERIALS AND METHODS

### A. Kather-CRC-Data set

This dataset contains non-overlapping 100,000 image patches, which include histological images of healthy tissue and CRC in humans (H&E). Each image is 224x224 pixels (px), with a pixel size of 0.5 microns (MPP). Adipose tissue, background (no tissue), detritus, lymphocytes, normal mucosa, mucus, stroma, muscle, and tumor epithelium were the nine types of tissues that were chosen from their database. The NCT Biobank and the UMM Pathology Archive provided the 86 formalin-fixed paraffin-embedded (FFPE) samples from which these images were manually retrieved. The tissue samples included CRC original tumor slides and tumor tissue from CRC liver metastases. To improve variety, non-tumorous gastrectomy specimen sections were included to the normal tissue classes.

Five samples of each CRC tissue type are shown in Fig. 2 from the Kather-CRC-NCT-CRC-HE-100K database. Tenfold cross validation was performed by J. N. Kather et al. [12] (<http://dx.doi.org/10.5281/zenodo.1214456>) to assess texturing approaches. The image composition of the databases NCT-CRC-HE-100K and CRC-VAL-HE-7K is shown in Table I.

TABLE I. DATABASE COMPOSITION

Class	Number of Images in NCT-CRC-HE-100K Database	Number of Images in CRC-VAL-HE-7K Database
adipose tissue	10,407	1,338
background (no tissue)	10,566	847
debris	10,512	339
lymphocytes	11,557	634
mucus	8,896	1,035
muscle	13,536	592
normal mucosa	8,763	741
stroma	10,446	421
tumor epithelium	14,317	1,233

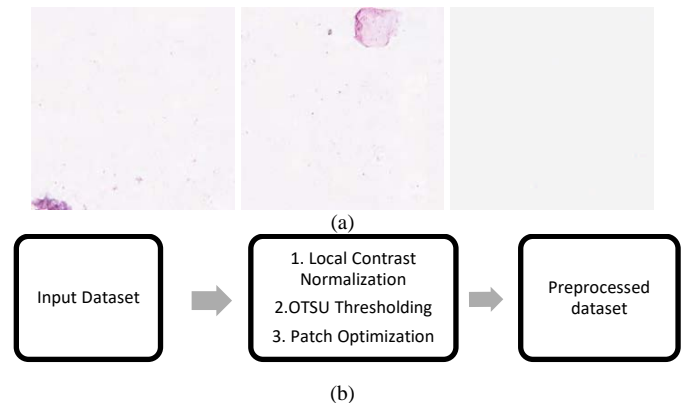


Fig. 1. Empty Patches in Database (b) Proposed Preprocessing Workflow.

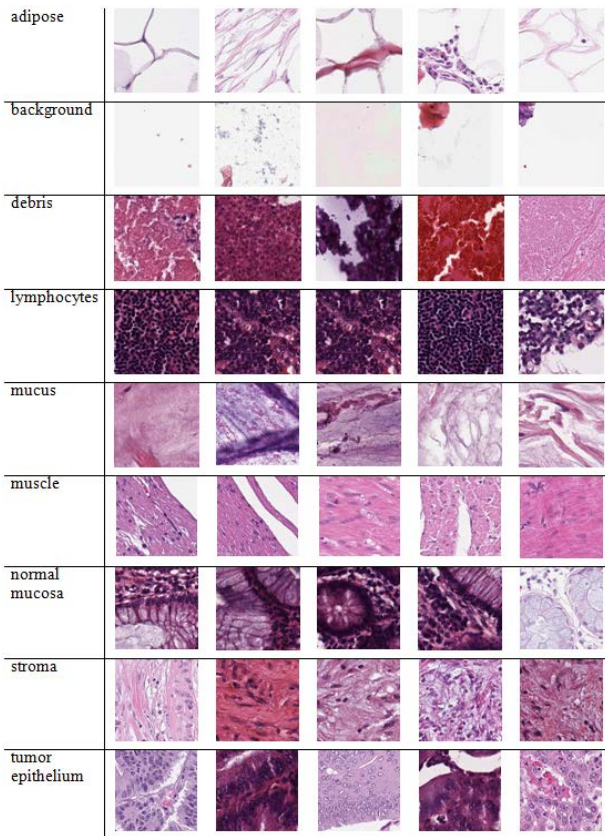


Fig. 2. Samples from the Kather-CRC- Database [12].

### B. Preprocessing of Dataset

By removing the empty tissue patch from the dataset, extra computations have been avoided on the non-tissue regions of the slide. There are many different techniques to evaluate an image's contrast. In deep learning, the standard deviation of an image's pixels or a region of an image is commonly referred to as contrast in equation (1) and (2).

$$\sqrt{\frac{1}{3rc} \sum_{i=1}^r \sum_{j=1}^c \sum_{k=1}^3 (X_{i,j,k} - \bar{X})^2} \quad (1)$$

where  $\bar{X}$  is the mean intensity of the entire image:

$$\bar{X} = \frac{1}{3rc} \sum_{i=1}^r \sum_{j=1}^c \sum_{k=1}^3 X_{i,j,k} \quad (2)$$

compute the LCN as proposed in [18] which is given in equation (3)

$$V_{i,j,k} = X_{i,j,k} - \sum_{i,p,q} w_{pq} X_{i,j+p,k+q} \quad (3)$$

Where  $w_{pq}$  is the Gaussian waiting window.

The RGB colour scheme of the low-resolution image was changed to LAB colour space before applying OTSU's threshold. After thresholding, binary morphological techniques were carried out to assist in the accurate patch extraction at small tissue regions and tissue borders. Fig. 1(a) shows few

samples of empty patches available in the database and Fig. 1(b) shows the proposed preprocessing flow for preprocessing the database. Even after separating the tissue region there is a chance of extracting patches with no information. So, one more step of patch optimization has been added to discard empty patches as shown in Fig. 1(b).

### C. CNN Architectures

Three of the most powerful fine-tuned CNN architectures, Alexnet, GoogLeNet, and Inception-v3 have been tested. Pre-trained models have been employed in this instance that was developed using the Kather-CRC-database [12].

#### D. Alexnet

The architecture is made up of eight layers: five convolutional layers and three fully connected layers. However, this is not what distinguishes AlexNet from other convolutional neural networks; rather, they are some of the characteristics that are employed. AlexNet uses Rectified Linear Units (ReLU) in place of the tanh function, which was referred as the industry standard. ReLU outperforms Tanh in terms of training velocity. A CNN utilizing ReLU was able to achieve a 25% error on the CIFAR-10 dataset six times quicker. CNNs frequently "pool" the outputs of neighboring neural groups without any overlap. However, after adding the overlapping, the error was reduced by roughly 0.5%, and it was shown that models with overlapping pooling are often more difficult to overfit. Overfitting was a serious concern for AlexNet.

#### E. GoogLeNet

The primary design of GoogLeNet [25] enhances computational capabilities inside the network model to encompass inception layers with the goal of minimizing complexity. By adding 1x1 convolutional layers to the network and using a different kernel, it not only enhances the depth but also the width of the architectural approach. In order to capture sparse correlation patterns, this lowers the number of computing levels

#### F. Inception-v3

The third iteration of the Inception networks family, which was initially introduced in [27], is known as Inception-v3 [34]. Inception block uses stacked 1x1 convolutions to reduce dimensionality, enabling fast computing and deeper networks. Unlike previous CNNs, which stacked kernel filter sizes sequentially, Inception architectures run several kernel filters with varying size on the same level. Making the networks larger rather than deeper is meant by this. The authors in [22,23] depicts the architecture of Inception-v3, which differs from the original Inception versions in a number of ways. These enhancements include propagating label information further down the network via an auxiliary classifier, factorized 7x7 convolutions, and label smoothing.

#### G. Fine-Tuning of Selected Models

Fine-tuning is a transfer learning concept in which information gained via training with one kind of difficulty is applied to training with another similar task or area [35]. The initial layers of deep learning algorithms are instructed to identify task-specific traits. The transfer learning phase is used

to remove few final layers of learnt network which can then be retrained with better task specified layers. Even if fine-tuned learning trials involve some learning, they nonetheless proceed far more quickly than learning from beginning [36]. Additionally, compared to models created from scratch, they are more accurate.

Data augmentation was used to fine-tune CNN Alexnet, Inceptionv3, GoogLeNet, and architecture using the Nct-Crc-He-100k and CRC-VAL-HE-7K datasets. The pretrained model has undergone the following adjustment.

1) The overfitting is greater if the size of the target data set is smaller and more comparable to the size of the training data set. The amount of overfitting that necessitates fine-tuning the data set for the pre-trained model is minimal if the target data set is bigger and comparable in size to the training data [37]. Therefore, a dropout layer has been added with probability 0.6 to the network to replace the final dropout layer, "pool5-drop 7x7 s1," which will randomly set certain features to zero.

2) Frequency can be modified based on the number of images allocated for training as follows.

$$\text{Validation Frequency} = \left\lfloor \frac{\text{Number of Images}}{\text{Batch Size}} \right\rfloor$$

3) The models were developed and then loaded with ImageNet pre-trained weights. As a result, a new fully-connected layer was developed in order to conduct the classification layer.

#### IV. EXPERIMENTAL RESULTS

##### A. Experimental Setup

The main components of the hardware environment are an Radeon RX 550X video card and an Intel Core i7-85650 CPU with 16 GB of RAM. Matlab R2020a is the software environment for language programming on a Windows 10 computer.

##### B. Dataset

To evaluate the fine-tuned pretrained model, two databases have been used: CRC-VAL-HE-7K and NCT-CRC-HE-100K. The bigger dataset of 100,000 non-overlapping image patches, NCT-CRC-HE-100K, and the smaller dataset of 7180, CRC-VAL-HE-7K, were chosen for testing. The comparison criteria used are displayed in Tables II and III. A total of 40% of the images in each dataset were used as the training set, 20% as the validation set, and 40% as the testing set.

##### C. Discussion

In the area of machine learning for image processing, deep learning models have prevailed. The possibility to extend the study and application to the identification and categorization of tumor epithelium in high resolution images is presented by

advancements in deep learning and image processing. However, the main problem with high quality images is training time.

TABLE II. PARAMETER VALUES TAKEN FOR COMPARISON (NCT-CRC-HE-100K DATABASE)

Fields	Size
Number of Classes	9
Dropout Probability	0.8 Vs pretrained models
Batch size	64
Epoch	15
Iterations	9360
Learning rate	1e <sup>-05</sup>

TABLE III. PARAMETER VALUES TAKEN FOR COMPARISON (CRC-VAL-HE-7K DATABASE)

Fields	Size
Number of Classes	9
Dropout Probability	0.4 Vs pretrained models
Batch size	64
Epoch	15
Iterations	660
Learning rate	1e <sup>-05</sup>

TABLE IV. TRAINING, VALIDATION AND TESTING ACCURACY OF NCT-CRC-HE-100K DATABASE

(Batch size = 64, Iterations = 9360, Learning rate = 1e<sup>-05</sup>, Epoch = 15)

Model	Training Acc. (%)	Validation Acc. (%)	Training Loss	Testing Acc. (%)	Testing Loss
Alexnet	95	95.03	0.42	94.5	0.39
GoogLeNet	95	94.08	0.38	94.3	0.29
InceptionV3	98	97.79	0.21	97.42	0.25

TABLE V. TRAINING, VALIDATION AND TESTING ACCURACY OF CRC-VAL-HE-7K DATABASE

(Batch size = 64, Iterations = 660, Learning rate = 1e<sup>-05</sup>, Epoch = 15)

Model	Training Acc. (%)	Validation Acc. (%)	Training Loss	Testing Acc. (%)	Testing Loss
Alexnet	93	93.18	0.51	93.2	0.43
GoogLeNet	92	91.36	0.59	91.4	0.7
InceptionV3	89.4	89.57	0.71	89	0.73

The deep learning architectures are adjusted in accordance with Section 3.G's instructions. Fig. 3 to 7 displays the experiment's findings. The accuracy and dropout probability of deep learning models (Alexnet, GoogLeNet and Inceptionv3) are shown in Fig. 3 and 4. For the pretrained model the default dropout parameter is 0.5 and it has been adjusted to 0.4 (for small dataset) and 0.8 (for large dataset) which is shown in Fig. 3 and 4. Therefore over fitting issue is properly handled using the dropout parameter which results good in both training set and the validation set. As the validation frequency has been generalized based on the size of the database, the batch loss and validation loss are very less which is shown in the Fig. 5, 6 and 7. Based on the above adjustments, all the models' accuracy increased after 15 epochs.

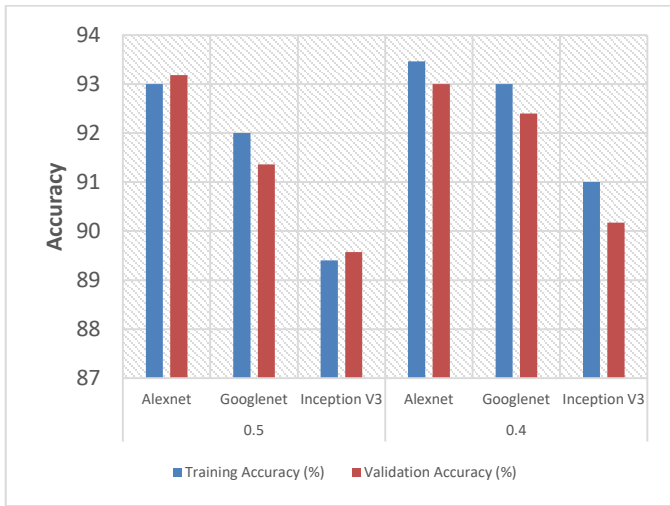


Fig. 3. Accuracy Comparison with and without Dropout Layer (CRC-VAL-HE-7K Database).

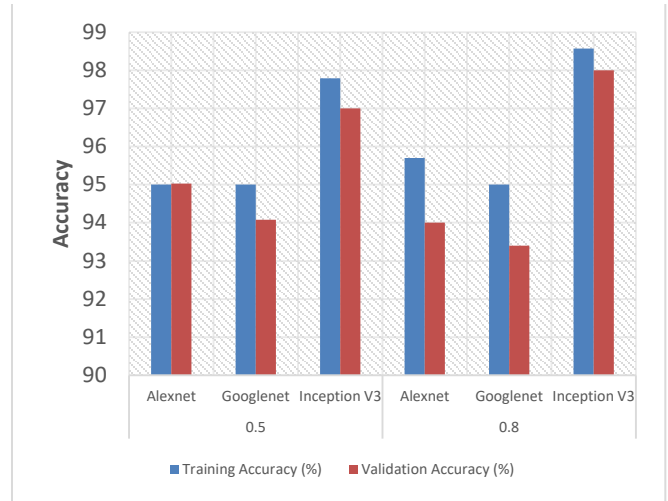


Fig. 4. Accuracy Comparison with and without Dropout Layer (NCT-CRC-HE-100K DATABASE).

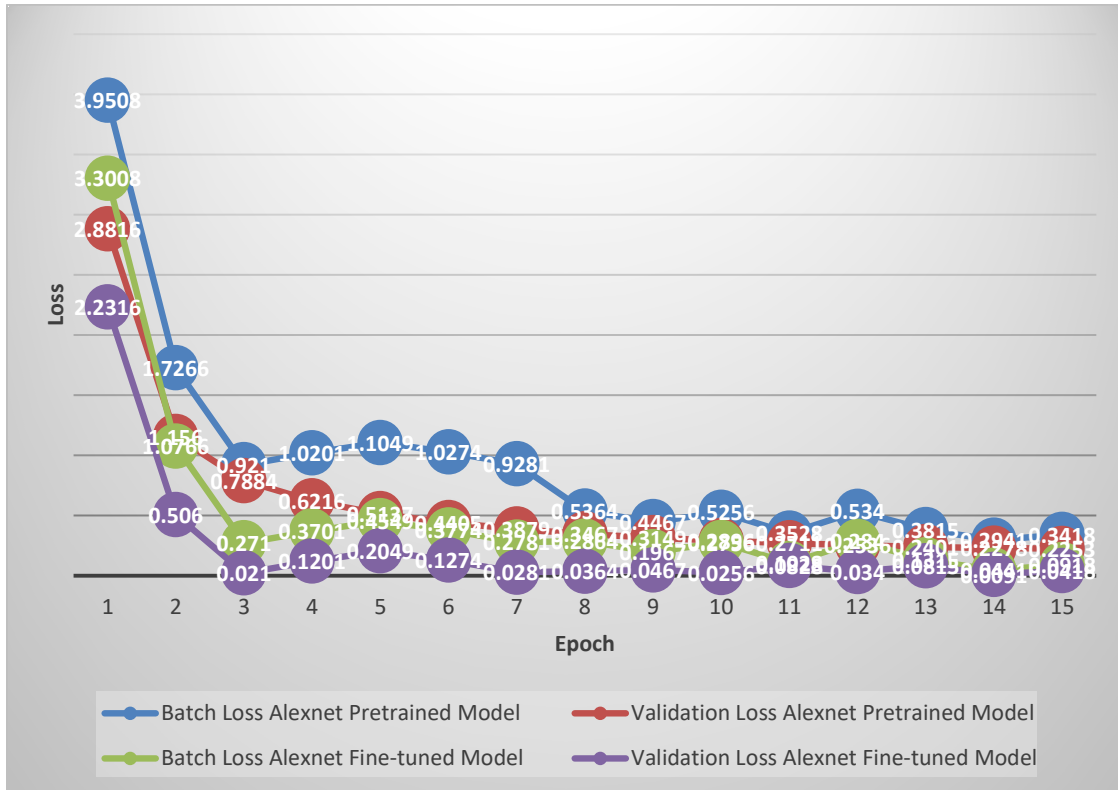


Fig. 5. Epoch Vs Batch Loss and Validation Loss of Alexnet.



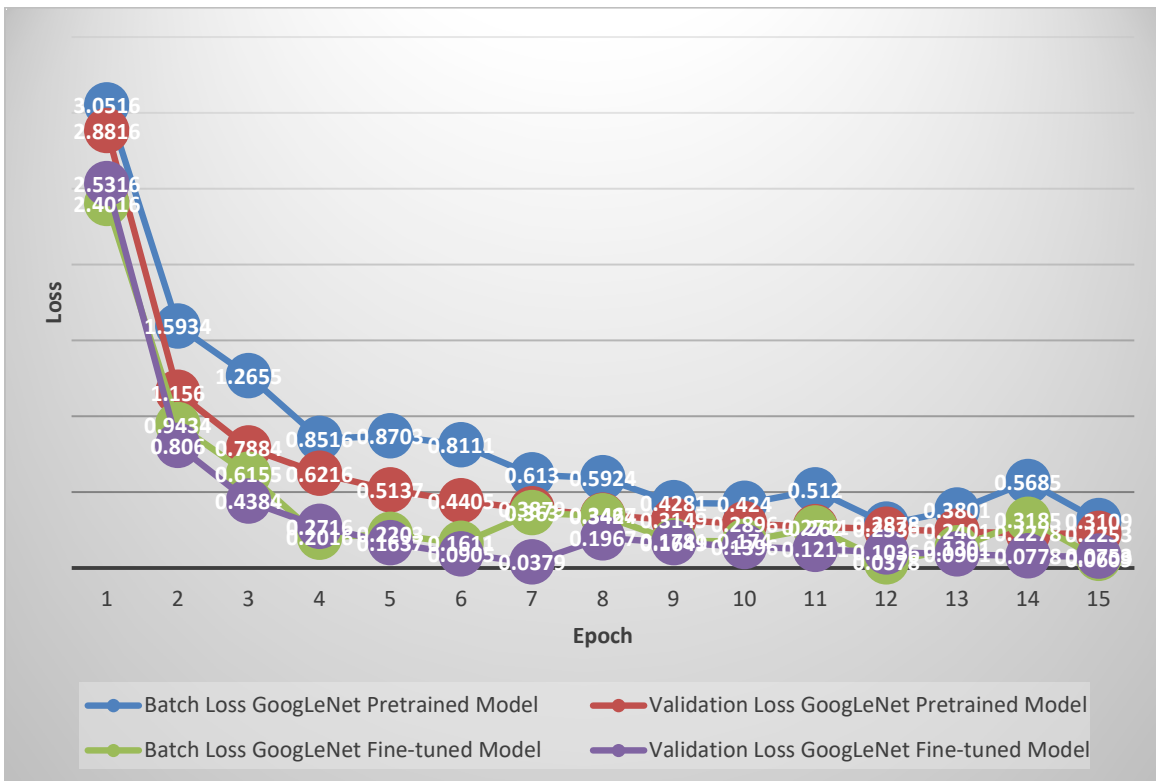


Fig. 6. Epoch Vs Batch Loss and Validation Loss of GoogLeNet.

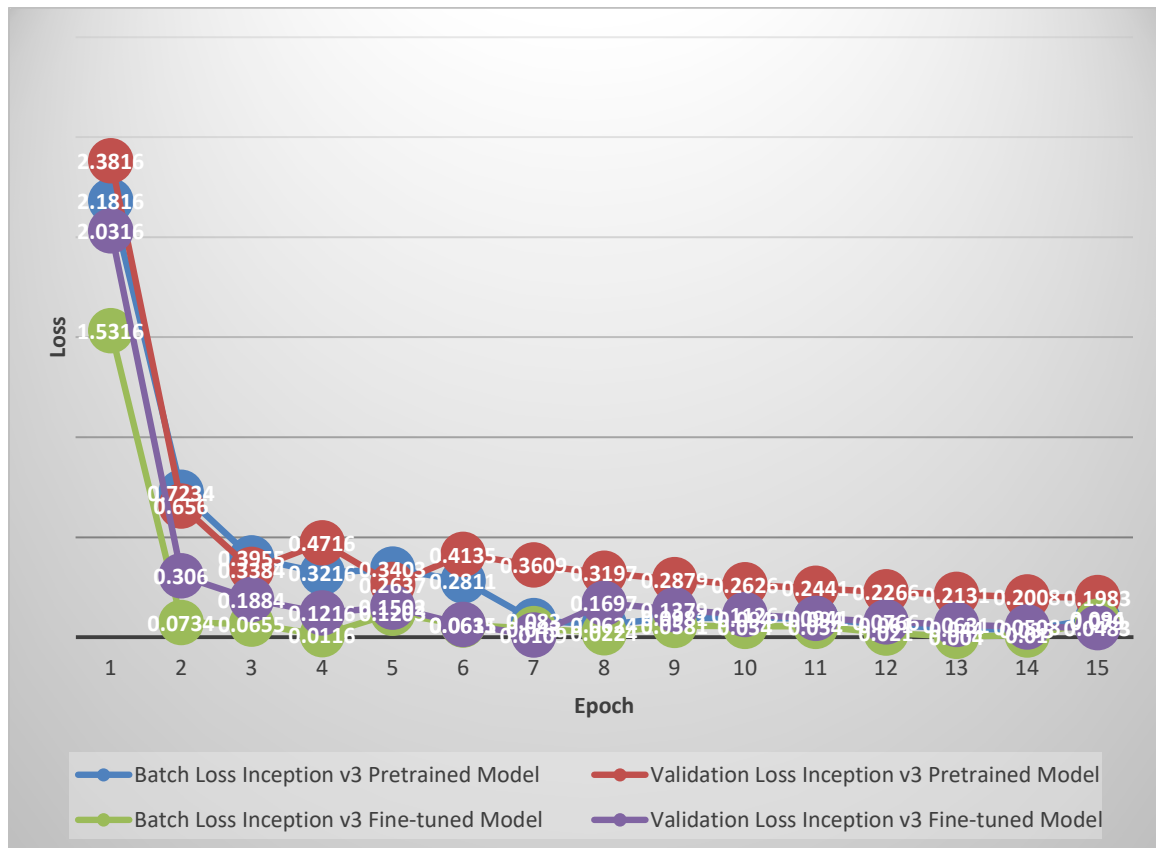


Fig. 7. Epoch Vs Batch Loss and Validation Loss of Inception v3.



TABLE VI. TISSUE AND TUMOR EPITHELIUM CLASSIFICATION ACCURACY USING PRETRAINED MODELS (CRC-VAL-HE-7K DATABASE)

Tissue Class	Alexnet	GoogLeNet	Inceptionv3
Adipose	97.8	97.6	94.23
Background	99.4	99.7	98.4
Debris	99.2	91.7	99
Lymphocytes	92.7	92.6	95.1
Mucus	94.6	92.7	85.4
Smooth Muscle	81	78.0	86.36
Normal Colon Mucosa	85.9	93.3	85.98
Stroma	84.7	88.2	90.26
Tumor epithelium	88.7	89.9	92.45

TABLE VII. TISSUE AND TUMOR EPITHELIUM CLASSIFICATION ACCURACY USING FINE-TUNED MODELS (CRC-VAL-HE-7K DATABASE)

Tissue Class	Alexnet	GoogLeNet	Inceptionv3
Adipose	98.5	98.1	95.73
Background	99.6	99.7	99.9
Debris	99.5	93.5	99.5
Lymphocytes	93.1	92.5	96.6
Mucus	95.2	93.4	86.9
Smooth Muscle	85.23	84.5	87.86
Normal Colon Mucosa	89.56	95.2	87.48
Stroma	90.21	91.78	92.76
Tumor epithelium	92.24	92.45	95.95

TABLE VIII. TISSUE AND TUMOR EPITHELIUM CLASSIFICATION ACCURACY USING PRETRAINED MODELS (NCT-CRC-HE-100K DATABASE)

Tissue Class	Alexnet	GoogLeNet	Inceptionv3
Adipose	97.5	98.31	99.61
Background	98.5	99.57	99.4
Debris	93.5	94.6	95.9
Lymphocytes	97.6	98.45	99.75
Mucus	95.4	96.21	97.51
Smooth Muscle	95.4	96.41	97.71
Normal Colon Mucosa	91.9	93.8	95.1
Stroma	88.4	91.3	92.6
Tumor epithelium	94.9	95.8	97.1

TABLE IX. TISSUE AND TUMOR EPITHELIUM CLASSIFICATION ACCURACY USING FINE-TUNED MODELS (NCT-CRC-HE-100K DATABASE)

Tissue Class	Alexnet	GoogLeNet	Inceptionv3
Adipose	98.4	99.09	99.5
Background	99.4	99.35	99.29
Debris	94.4	95.38	96.79
Lymphocytes	98.5	99.23	99.64

Mucus	96.3	96.99	98.4
Smooth Muscle	96.3	97.19	98.6
Normal Colon Mucosa	92.8	94.58	95.99
Stroma	89.3	92.08	93.49
Tumor epithelium	95.8	96.58	97.99

TABLE X. TRAINING TIME FOR FINE-TUNED MODELS

Models	Training Time CRC-VAL-HE-7K Database in Seconds		Training Time CRC-VAL-HE-7K Database in Seconds	
	No preprocessing	Proposed preprocessing	No preprocessing	Proposed preprocessing
Alexnet	3060	2564	307380	256897
GoogLeNet	8520	6295	83880	51520
Inception V3	41520	11265	110735	92545

The validation accuracy of the proposed finetuned Alexnet is better when compared to the validation accuracy proposed in [38]. The later achieved 91.8% accuracy with 25 epoch whereas this work achieved 93.18 % (Table V) and 95.03 (Table IV) in 15 epochs using the same database. The author in [39] claimed, that the network they developed, SCDNet, achieved an accuracy of 96.91% which is 4% more than the pretrained inception v3 model. The proposed fine-tuned inceptionv3 model achieves 97.79% accuracy which is better when compared to SCDNet. Additionally, as shown in Fig. 5 to 7, good accuracy results were obtained for all the chosen models even after the 30th training iteration with much reduced batch and validation loss.

GoogLeNet and Inceptionv3 models regularly outperform Alexnet among the three models chosen. Tables VI, VII, VIII, and IX shows the accuracy of nine tissue classes mentioned in Table I. As shown in Tables VI, VII, VIII, and IX, the accuracy of the tumor epithelium classification is higher in inceptionv3. Table X shows the training time required to execute the finetuned models based on the experimental setup. As indicated in Table X, training time has been reduced by at least 30% using the suggested preprocessing workflow. Overall, Alexnet fared badly, having the least accuracy and having the largest batch and validation loss, whereas finetuned inceptionv3 performed well, having the highest accuracy and the lowest batch and validation loss.

## V. CONCLUSION

A workflow has been suggested to preprocess the dataset in this article. Additionally, the most advanced deep convolutional neural network for classifying tissues and tumor epithelium is being tuned and evaluated. The architectures under consideration are Inceptionv3, GoogLeNet, and Alexnet. According to the experiment, Inceptionv3 tends to provide a cogent accuracy increase with increasing epochs, without showing any signs of overfitting or performance degradation. Additionally, Inceptionv3 performs better in classification exhibitions with few parameters and fair processing time. Inceptionv3 outperforms the other architectures with a test accuracy score of 97.42% for the 15 epoch. Thus, Inceptionv3

is a promising design for the goal of identifying tumor epithelium. The proposed parameters can be extended with other pretrained models and the performance can be compared with the parameters F1-score, AUC, Recall and Precision. Also, a new generalized Deep CNN model can be designed which satisfies the proposed adjustment parameters.

#### REFERENCES

- [1] Farahani, N.; Parwani, A.V.; Pantanowitz, L. Whole slide imaging in pathology: Advantages, limitations, and emerging perspectives. *Pathol. Lab. Med. Int.* 2015, 7, 4321.
- [2] M. Zeeshan, B. Salam, Q. S. B. Khalid, S. Alam, and R. Sayani, "Diagnostic accuracy of digital mammography in the detection of breast cancer." *J. Cureus*, vol. 10, no. 4, p. e2448, Apr. 2018, doi: 10.7759/cureus.2448.
- [3] M. N. Gurcan, L. E. Boucheron, A. Can, A. Madabhushi, N. M. Rajpoot, and B. Yener, "Histopathological image analysis: A review," *IEEE Rev. Biomed. Eng.*, vol. 2, pp. 147–171, 2009, doi: 10.1109/RBME.2009.2034865.
- [4] J. Hipp, A. Fernandez, C. Compton, and U. Balis, "Why a pathology image should not be considered as a radiology image," *J. Pathol. Informat.*, vol. 2, no. 1, p. 26, 2011, doi: 10.4103/2153-3539.82051.
- [5] M. D. Pickles, P. Gibbs, A. Hubbard, A. Rahman, J. Wiczorek, and L. W. Turnbull, "Comparison of 3.0T magnetic resonance imaging and X-ray mammography in the measurement of ductal carcinoma in situ: A comparison with histopathology," *Eur. J. Radiol.*, vol. 84, no. 4, pp. 603–610, Apr. 2015, doi: 10.1016/j.ejrad.2014.12.016.
- [6] Pantanowitz, L.; Sharma, A.; Carter, A.B.; Kurc, T.; Sussman, A.; Saltz, J. Twenty years of digital pathology: An overview of the road travelled, what is on the horizon, and the emergence of vendor-neutral archives. *J. Pathol. Inform.* 2018, 9, 40.
- [7] Egeblad, M.; Nakasone, E.S.; Werb, Z. Tumors as organs: Complex tissues that interface with the entire organism. *Dev. Cell* 2010, 18, 884–901.
- [8] Huijbers, A.; Tollenaar, R.; Pelt, G.W.; Zeestraten, E.C.M.; Dutton, S.; McConkey, C.C.; Domingo, E.; Smit, V.; Midgley, R.; Warren, B.F. The proportion of tumor-stroma as a strong prognosticator for stage II and III colon cancer patients: Validation in the VICTOR trial. *Ann. Oncol.* 2013, 24, 179–185.
- [9] Marusyk, A.; Almendro, V.; Polyak, K. Intra-tumor heterogeneity: A looking glass for cancer? *Nat. Rev. Cancer* 2012, 12, 323–334.
- [10] Bray, F.; Ferlay, J.; Soerjomataram, I.; Siegel, R.L.; Torre, L.A.; Jemal, A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* 2018, 68, 394–424.
- [11] 7. Sirinukunwattana, K.; Snead, D.; Epstein, D.; Aftab, Z.; Mujeeb, I.; Tsang, Y.W.; Cree, I.; Rajpoot, N. Novel digital signatures of tissue phenotypes for predicting distant metastasis in colorectal cancer. *Sci. Rep.* 2018, 8, 1–13.
- [12] Kather, Jakob Nikolas, et al. "Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study." *PLoS medicine* 16.1 (2019): e1002730..
- [13] 9. Nearchou, I.P.; Soutar, D.A.; Ueno, H.; Harrison, D.J.; Arandjelovic, O.; Caie, P.D. A comparison of methods for studying the tumor microenvironment's spatial heterogeneity in digital pathology specimens. *J. Pathol. Inform.* 2021, 12, 6.
- [14] Sigirci, I. Onur, Abdulkadir Albayrak, and Gokhan Bilgin. "Detection of mitotic cells in breast cancer histopathological images using deep versus handcrafted features." *Multimedia Tools and Applications* 81.10 (2022): 13179-13202.
- [15] Huynh BQ, Li H, and Giger ML. Digital mammographic tumor classification using transfer learning from deep convolutional neural networks. *J. Med. Imaging* 3:34501, 2016.
- [16] Tran H, Phan H, Kumar A, Kim J, and Feng D. Transfer Learning of a Convolutional Neural Network for Hep-2 Cell Image Classification 2012:1208–1211, 2016.
- [17] Zhang R, Zheng Y, Mak TWC, Yu R, Wong SH, Lau JYW, and Poon CCY. Automatic Detection and Classification of Colorectal Polyps by Transferring Low-Level CNN Features from Nonmedical Domain. *IEEE J. Biomed. Heal. Informatics* 21:41–47, 2017.
- [18] Shin HC, Roth HR, Gao M, Lu L, Xu Z, Nogues I, Yao J, Mollura D, and Summers RM. Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning. *IEEE Trans. Med. Imaging* 35:1285–1298, 2016
- [19] Pan Y, Huang W, Lin Z, Zhu W, Zhou J, Wong J, and Ding Z. Brain Tumor Grading Based on Neural Networks and Convolutional Neural Networks. *Eng. Med. Biol. Soc. (EMBC), 2015 37th Annu. Int. Conf. IEEE* 699–702, 2015. doi:10.1109/EMBC.2015.7318458
- [20] Kather, J.N.; Weis, C.A.; Bianconi, F.; Melchers, S.M.; Schad, L.R.; Gaiser, T.; Marx, A.; Zöllner, F.G. Multi-class texture analysis in colorectal cancer histology. *Sci. Rep.* 2016, 6, 27988
- [21] Kothari, S.; Phan, J.H.; Young, A.N.; Wang, M.D. Histological image classification using biologically interpretable shape-based features. *BMC Med. Imaging* 2013, 13, 9.
- [22] Javed, S.; Mahmood, A.; Fraz, M.M.; Koohbanani, N.A.; Benes, K.; Tsang, Y.W.; Hewitt, K.; Epstein, D.; Snead, D.; Rajpoot, N. Cellular community detection for tissue phenotyping in colorectal cancer histology images. *Med. Image Anal.* 2020, 63, 101696.
- [23] Bejnordi, B.E.; Mullooly, M.; Pfeiffer, R.M.; Fan, S.; Vacek, P.M.; Weaver, D.L.; Herschorn, S.; Brinton, L.A.; van Ginneken, B.; Karssemeijer, N. Using deep convolutional neural networks to identify and classify tumor-associated stroma in diagnostic breast biopsies. *Mod. Pathol.* 2018, 31, 1502–1512.
- [24] Nanni, L.; Brahnam, S.; Ghidoni, S.; Lumini, A. Bioimage classification with handcrafted and learned features. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 2018, 16, 874–885.
- [25] Bougourzi, F.; Dornaika, F.; Mokrani, K.; Taleb-Ahmed, A.; Ruichek, Y. Fusion Transformed Deep and Shallow features (FTDS) for Image-Based Facial Expression Recognition. *Expert Syst. Appl.* 2020, 156, 113459.
- [26] Wang, S.; Yang, D.M.; Rong, R.; Zhan, X.; Fujimoto, J.; Liu, H.; Minna, J.; Wistuba, I.I.; Xie, Y.; Xiao, G. Artificial intelligence in lung cancer pathology image analysis. *Cancers* 2019, 11, 1673.
- [27] Ouahabi, A.; Taleb-Ahmed, A. Deep learning for real-time semantic segmentation: Application in ultrasound imaging. *Pattern Recognit. Lett.* 2021, 144, 2–34.
- [28] Cascianelli, S.; Bello-Cerezo, R.; Bianconi, F.; Fravolini, M.L.; Belal, M.; Palumbo, B.; Kather, J.N. Dimensionality reduction strategies for cnn-based classification of histopathological images. In *Proceedings of the International Conference on Intelligent Interactive Multimedia Systems and Services, Gold Coast, Australia, 20–22 May 2018*; pp. 21–30.
- [29] Kather, J.N.; Krisam, J.; Charoentong, P.; Luedde, T.; Herpel, E.; Weis, C.A.; Gaiser, T.; Marx, A.; Valous, N.A.; Ferber, D. Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study. *PLoS Med.* 2019, 16, e1002730.
- [30] Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* 2014, arXiv:1409.1556.
- [31] Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Commun. ACM* 2017, 60, 84–90.
- [32] Iandola, F.N.; Han, S.; Moskewicz, M.W.; Ashraf, K.; Dally, W.J.; Keutzer, K. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and
- [33] Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015*; pp. 1–9.
- [34] He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016*; pp. 770–778.
- [35] Pan, S.J., Fellow, Q.Y., 2009. A Survey on Transfer Learning, pp. 1–15.
- [36] Mohanty, S. P., D. P. Hughes, and M. Salathé. "Using deep learning for image-based plant disease detection 7, 1–10." (2016).

- [37] Senan, Ebrahim Mohammed, et al. "Classification of histopathological images for early detection of breast cancer using deep learning." *Journal of Applied Science and Engineering* 24.3 (2021): 323-329.
- [38] Izzaty, Al Mira Khonsa, et al. "Multiclass classification of histology on colorectal cancer using deep learning." *Commun. Math. Biol. Neurosci.* 2022 (2022)
- [39] Naeem, Ahmad, et al. "SCDNet: A Deep Learning-Based Framework for the Multiclassification of Skin Cancer Using Dermoscopy Images." *Sensors* 22.15 (2022): 5652.

# Predicting University Student Retention using Artificial Intelligence

Samer M. Arqawi<sup>1</sup>

Associate Professor at Industrial Management Department  
Palestine Technical University-Kadoorie, Palestine

Eman Akef Zitawi<sup>2</sup>

Researcher in Educational Administration  
College of Graduate Studies - Department of Educational  
Administration, Arab American University Palestine

Anees Husni Rabaya<sup>3</sup>

Head of the Health Administration Department  
Al Quds Open University

Basem S. Abunasser<sup>4</sup>

University Malaysia of Computer Science & Engineering  
(UNIMY), Cyberjaya, Malaysia

Samy S. Abu-Naser<sup>5</sup>

Professor of Data Science  
Faculty of Engineering and Information Technology, Al-Azhar University, Gaza, Palestine

**Abstract**—Based on the advancement in the field of Artificial Intelligence, there is still a room for enhancement of student university retention. The main objective of this study is to assess the probability of using Artificial Intelligence techniques such as deep and machine learning procedures to predict university student retention. In this study a variable assessment is carried out on the dataset which was collected from Kaggle repository. The performance of twenty supervised algorithms of machine learning and one algorithm of deep learning is assessed. All algorithms were trained using 10 variables from 1100 records of former university student registrations that have been registered in the University. The top performing algorithm after hyper-parameters tuning was NuSVC Classifier. Therefore, we were able to use the current dataset to create supervised Machine Learning (ML) and Deep Learning (DL) models for predicting student retention with F1-score (90.32 percent) for ML and the proposed DL algorithm with F1-score (93.05 percent).

**Keywords**—Artificial intelligence; machine learning; deep learning; retention; student; prediction

## I. INTRODUCTION

To begin with, Neural Networks remained extensively recognized as Artificial Neural Networks. A construction similar to the biological neurons implementation, with a learning configuration grounded on probability that gets input data to make a decision on an output. Texts, images and audio files are examples of input data. Deep Learning (DL) employs artificial neurons to work on high dimensional data. DL can accomplish tasks including information processing and communication patterns [1, 2, 3].

Fig. 1 exemplifies the rudimentary structure of Neural Networks, beginning with the first layer called the input and the last one is the output with the middle layers as the hidden layers therefore the name “deep”. With additional research, and with the implementation of backpropagation, the structure was enhanced. The algorithm of backpropagation has two

phases, the first one is for the input to forward the data in the direction of the output, the second one is to assess and approximate the error and back-propagating error values to the neural networks for correcting the error.

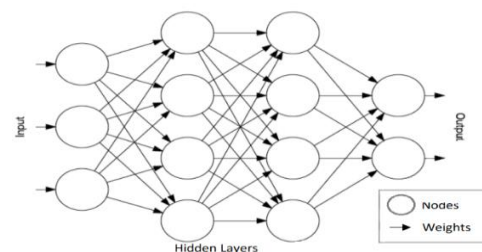


Fig. 1. Structure of Neural Networks

Every node in is an attribute pulled out from the portion beforehand. The count of neurons in a layer differs for every system, as c of the neuron in every layer should be capable of capturing the vital layer. The first layer of the model is usually high in terms of the count of neurons [4, 5].

DL defines an entirely new area of research with the architecture that has solved many of the problems that existed in the traditional systems in which the model was used for definite usage or exact commands [6].

DL was capable of a more unified model that could be useful on many applications in addition to diverse users. The structure of DL can be applied on numerous applications, like Google Assistant which can be implemented in mobile devices (Android), moreover can be implemented in search engines (Google) in addition to being applied on home devices (Google Home), and likewise employed in video captioning like video captioning on the YouTube [7, 8].

In this paper, some data classification techniques are applied to the evaluation of the dataset to predict whether students will continue to attend university or not by finding

the best classification technique in accordance with these measures: F1-score, Accuracy, Precision, Recall, and time performance.

We used 20 machine learning algorithms to predict whether students will continue to attend university or not. Furthermore, a deep learning model was proposed for the same purpose.

The aim of this study is to answer the following questions:

- Can we use the dataset at hand to create a supervised machine learning classifier to predict reliably whether a student will go to university or not?
- Can we use the dataset at hand to create a supervised Deep learning classifier to predict reliably whether a student will go to university or not?
- Would the performance of Deep Learning techniques be better than the machine learning techniques in this case?

## II. MACHINE LEARNING ALGORITHMS

Machine learning algorithms can be applied to solve multiple problems. Classification can be used to assign a category to items or answer “yes” or “no” questions. An example of a classifier is a program that categorizes news articles or a spam filter that answers the question “spam” or “not spam”. Regression algorithms can be used to predict values for items such as housing prices. Ranking can be used to order items according to a criterion and clustering can be used to partition items into homogeneous regions [9, 10, 11, 12, 13].

A decision tree is an example of a machine learning algorithm. Decision trees contain a root with a question. The root has branches to more nodes with questions and each path of nodes and branches ends with a leaf containing a label. The tree is traversed by taking decisions at each node based on the question until a leaf is found and the label for the leaf is returned as the answer to the root question [14, 15, 16].

Ensemble learning methods are boosting algorithms that combine several weak learners into a single strong learner. One example of a boosting algorithm, AdaBoost [17], builds this strong learner from several weak learners by making the

next learner focus on improving the mistakes of the previous learner. The learners are assigned weights by the algorithm after each iteration and the weights are used to merge the weak learners into a strong learner [18, 19].

Linear models predict values by seeking a hypothesis of a hyperplane that has the smallest outcome of an error function. The linear model has to predict the value of  $y$  for the input value  $x$ . Using a set of known inputs  $x$  and labeled output values  $y$  a model is evaluated which has the smallest error, in this case distance of the line from the points in the graph. The model with the smallest error is represented as a line passing through the points. New input values  $x$  can then have their output values  $y$  predicted using the model by inspecting where the line meets the input value  $x$  on the  $y$  axis. Linear models such as Logistic Regression can also be used to solve classification problems [20, 21, 22, 23, 24].

Support Vector Machines (SVM) can be applied to both regression and classification problems. SVMs seek to find a hyperplane which has the maximum margin from all inputs in the training set. SVMs differentiate and improve upon Linear Models with a better error or loss function. SVMs can be even more improved upon with the use of kernels to define non-linear decision boundaries [25, 26, 27, 28].

In addition to the aforementioned algorithms, a great deal of other algorithms exist. K-Nearest-Neighbor considers neighboring objects in the dataset to train a better model [29]. Naive Bayes (NB) algorithms use Bayes’ theorem to compute conditional probabilities [30]. Discriminant Analysis classifies objects in a dataset by identifying the best feature to discriminate between classes [31, 32].

## III. METHODOLOGY

In this section we will give details of the data collection, preparation, feature analysis, data splitting, modeling the proposed deep learning algorithm, training, validating and testing all the algorithms used in this study(as shown in Fig. 22).

### A. Dataset

The dataset was collected from Kaggle depository. The dataset consists of 11 features and contains 1000 records. Table I lists all the features available from the dataset.

TABLE I. A LIST OF AVAILABLE FEATURES

Feature	Type	Description
Type-school	object	Academic or Vocational
School-accreditation	object	A or B
Interest	object	Very Interested, Uncertain, Less Interested, Quiet Interested, Not Interested
Average-grades	numeric	75% to 98%
Gender	object	Male or Female
Residence	object	Urban, Rural
Parent-age	numeric	40 to 65 years
Parent-salary	numeric	1,000,000 to 10,000,000
House-area	numeric	20..120
Parent-was-in-university	Boolean	True or False
In-university	Boolean	True or False

B. Feature Analysis

1) *Correlation matrix*: The highest in negative correlation will be parents-was-in-university – parent-age, and with positive is parent-salary – in-university, house-area – in-university and average-grades – in-university as illustrated in Fig. 2.

2) *Histogram for distribution*: Fig. 3 shows the distribution of the features: parent-age, parent-salary, house-area, and average-grades.

3) *Density plot for distribution*: Fig. 4 illustrates the density plot distribution for parent-age, parent-salary, house-area, and average-grades. The overall distribution looks a little bit normal.

4) *Outliers*: Fig. 5 does not show so many outliers, so my decision is to keep it without further cleaning.

5) *Residence count based on university status*: Overall the Urban is more likely to go to university and the opposite happens in Rural as shown in Fig. 6.

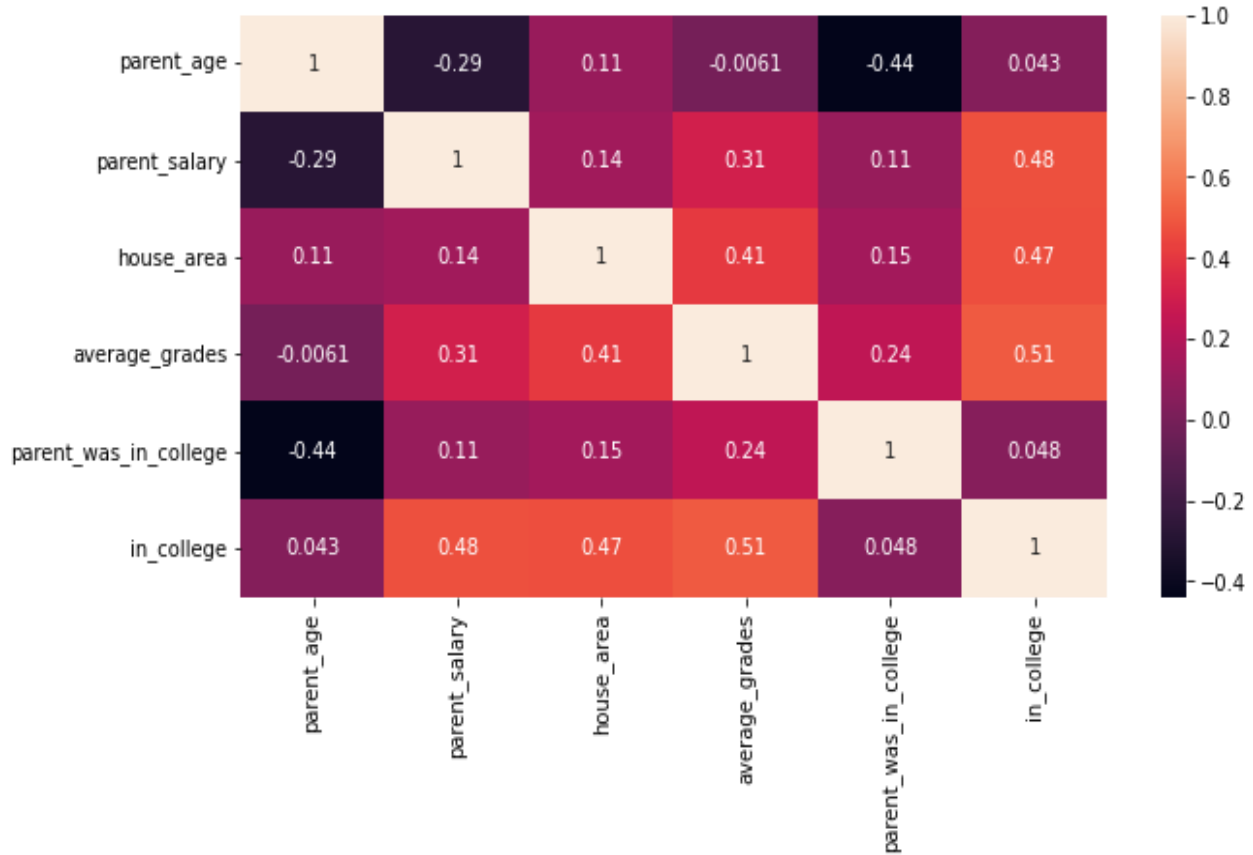


Fig. 2. Correlation Matrix.

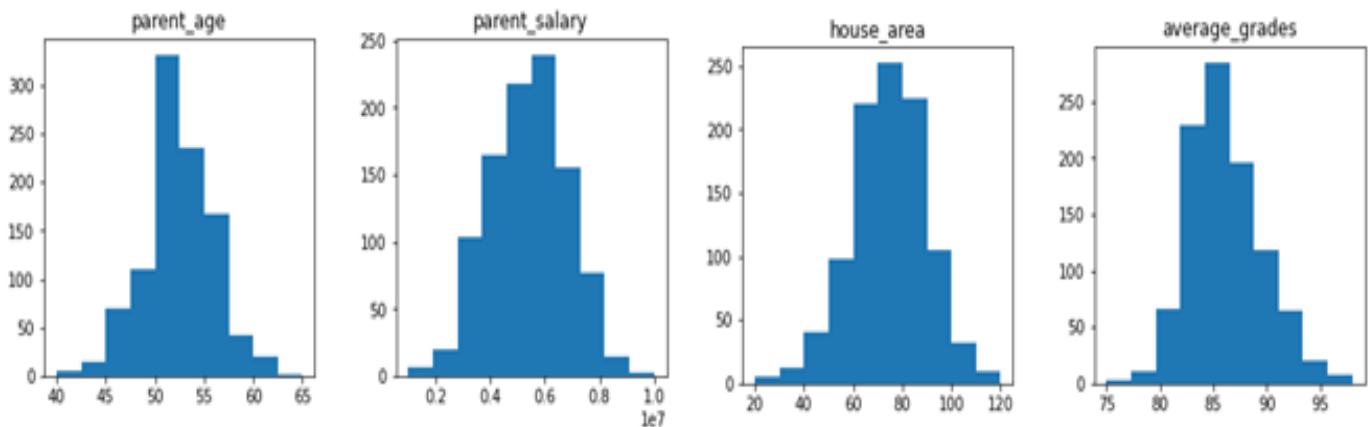


Fig. 3. Histogram for Distribution.



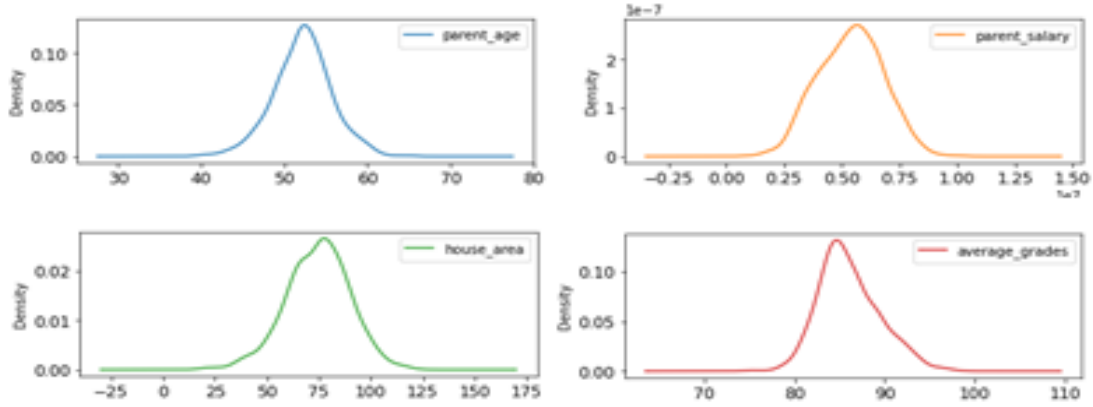


Fig. 4. Density Plot for Distribution.

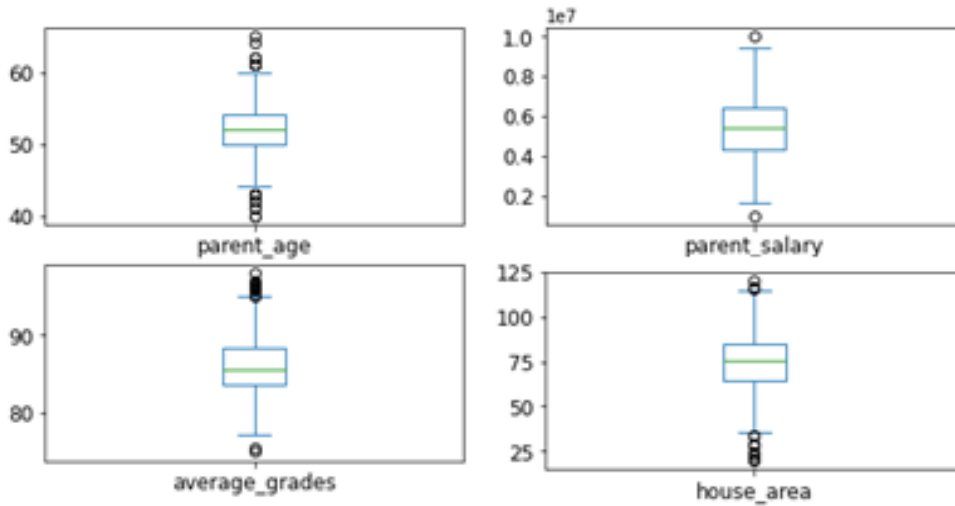


Fig. 5. Outliers of the Parent-Age, Parent-Salary, Average Grades, and House-Area.

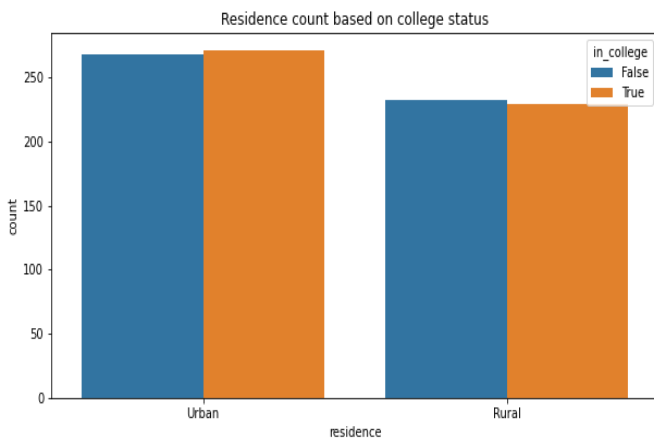


Fig. 6. Residence Count based on University Status.

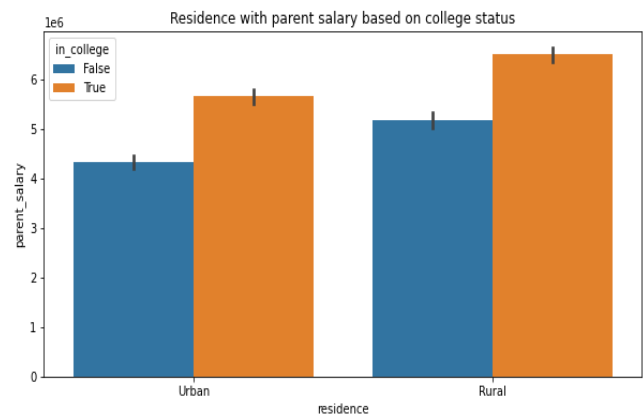


Fig. 7. Residence with Parent Salary based on University Status.

6) *Residence with parent salary based on university status*: Fig. 7 illustrates that the higher the salary the more likely to go to university for both Urban and Rural.

7) *Interest count based on university status*: From Fig. 8, Very Interested and Uncertain Interest are the two most categories attending the university, and Less Interested is the most in the category who does not attend the university.

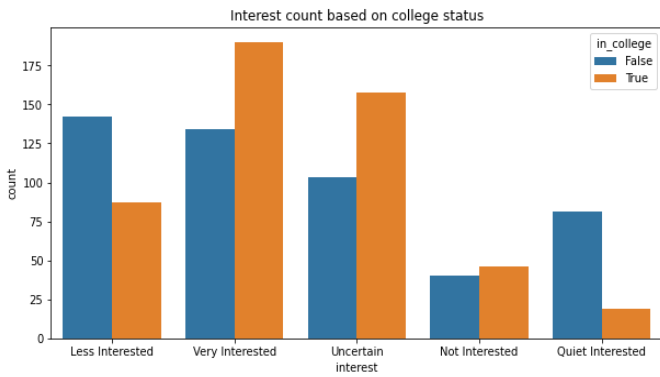


Fig. 8. Interest Count based on University Status.

8) *School type count based on university status:* Fig. 9 show that the Academic is the most in the category who attended the university.

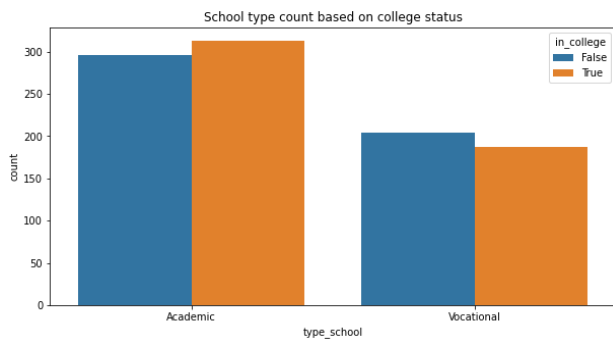


Fig. 9. School Type Count based on University Status.

9) *Gender count based on university status:* From Fig. 10, females are slightly higher to attend the university than Males.

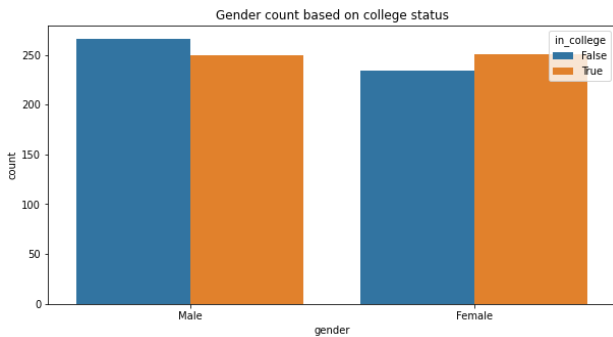


Fig. 10. Gender Count based on University Status.

10) *Parent age, residence and parent-was-in-university:* Fig. 11 illustrates that urban parents was slightly higher to attend the university than rural parents.

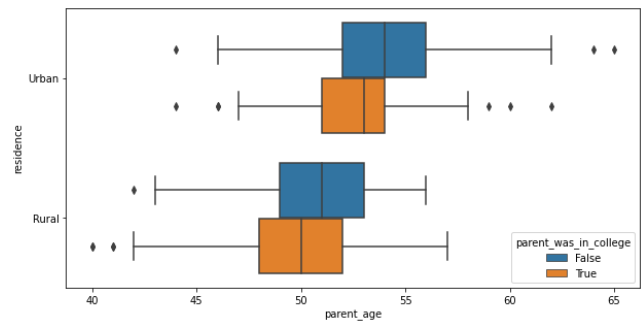


Fig. 11. Parent Age, Residence and Parent-was-in-University Relation.

11) *Parent salary, residence and parent-was-in-university:* Fig. 12 show that rural parents were higher salary and not attended to university and the opposite for urban parents.

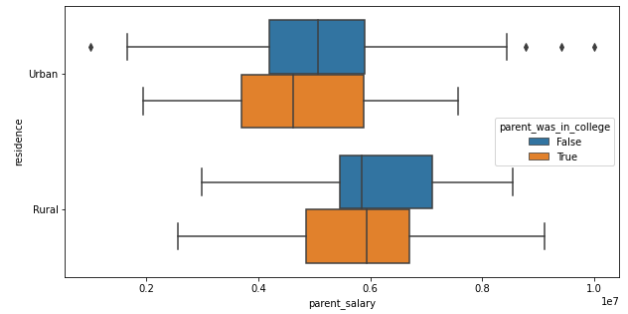


Fig. 12. Parent Salary, Residence and Parent-was-in-University.

12) *Parent salary, residence and in-university:* The more salary in both Rural and Urban parents the more likely for their child to attend the university as can be seen in Fig. 13.

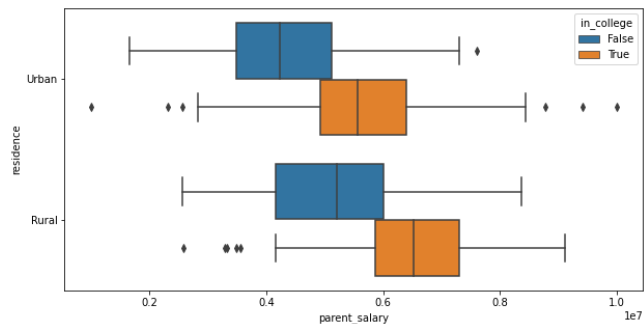


Fig. 13. Parent Salary, Residence and in-University.

13) *House-area, residence and in-university:* The more house area in both Rural and Urban the more likely for a child to attend the university, that's means by logic the family has more salary overall as can be seen in Fig. 14.

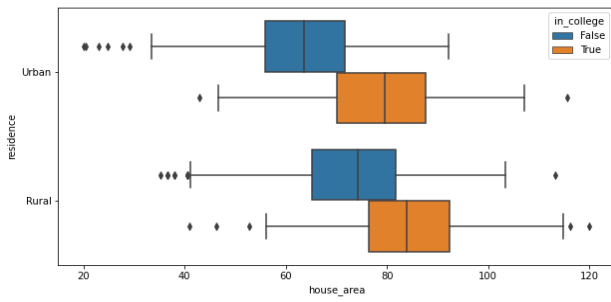


Fig. 14. House-Area, Residence and in-University.

14) *Average-grades, gender and in-university*: The more average grades in both males and females the more likely to attend university as shown in Fig. 15.

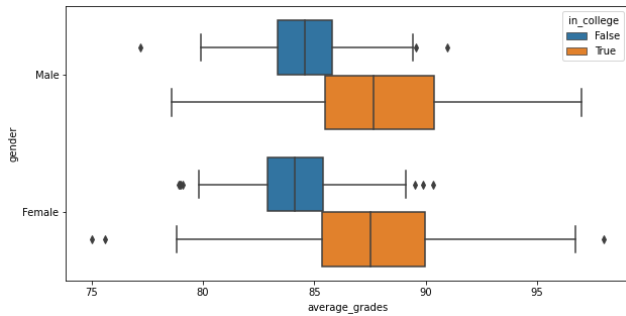


Fig. 15. Average-Grades, Gender and in-University.

15) *Average-grades and in-university*: The higher on average grades the more likely to attend the university as in Fig. 16.

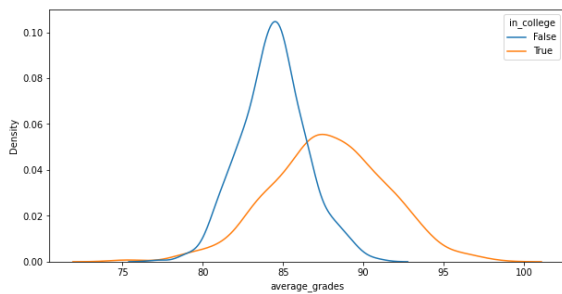


Fig. 16. Average-Grades and in-University.

16) *Parent-age and in-university*: Not so much effect by the age of parents for a child to attend the university as in Fig. 17.

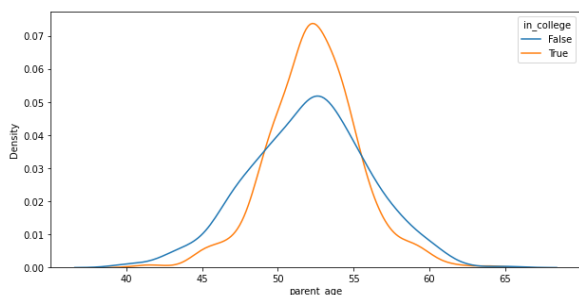


Fig. 17. Parent-Age and in-University.

17) *Parent- salary and in-university*: From Fig. 18, the higher on a parent's salary the more likely for the child to attend university.

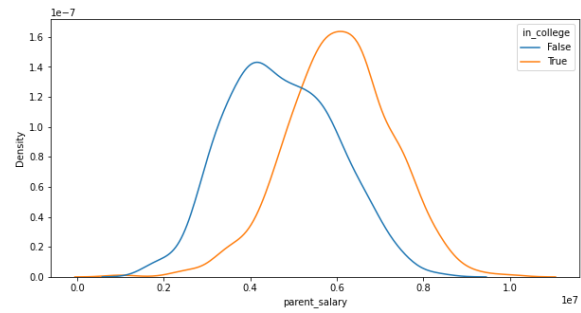


Fig. 18. Parent- Salary and in-University.

18) *House-area and in-university*: From Fig. 19, it can be seen that the higher the house area the more likely for the child to attend university.

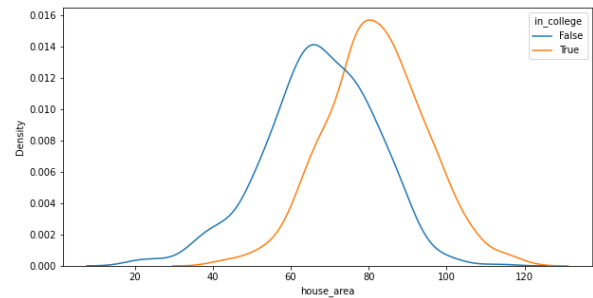


Fig. 19. House-Area and in-University.

19) *Converting categorical column to numeric ones*: We have converted the categorical type features (type-school, school-accreditation, gender, interest, residence, parent-was-in-university, and in-university) to numeric values.

20) *Class (in-university) Distributions*: We checked whether the class (in-university) is balanced or not. It turned out to be balanced as shown in Fig. 20.

C. Dataset Splitting

We have split the current dataset into three datasets: Training, validating, and testing datasets. The ratio of splitting was (80%, 10%, 10%).

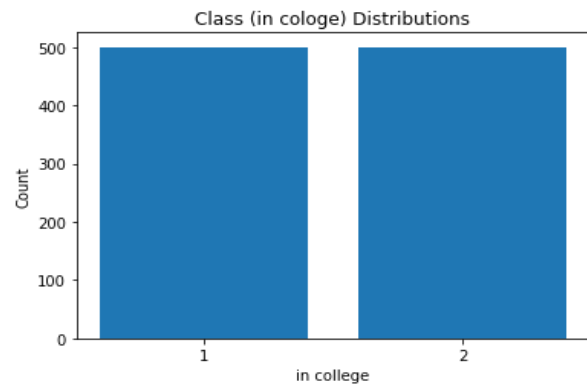


Fig. 20. Class (in-University) Distribution.

D. Description of the Algorithms used in this Study

There are many algorithms of Machine Learning that can be applied for the prediction of whether the student continues to attend university or not. We have trained, validated and tested 20 various ML algorithms on our current dataset. The algorithms that were used for prediction and analysis are from different categories of machine learning algorithms to predict whether students will continue to attend university or not.

Furthermore, a deep learning model was proposed to predict whether students will continue to attend university or not. The proposed DL model consists of 6 Dense layers: one input layer (10 features), 4 hidden layers (128, 64, 32, and 16 neurons), and one output layer with 2 classes and softmax function as shown in Fig. 21.

```

Model: "model_1"
-----
Layer (type)                Output Shape                Param #
-----
input_2 (InputLayer)        [(None, 10)]                0
dense_5 (Dense)              (None, 128)                 1408
dense_6 (Dense)              (None, 64)                  8256
dense_7 (Dense)              (None, 32)                  2080
dense_8 (Dense)              (None, 16)                  528
dense_9 (Dense)              (None, 2)                   34
-----
Total params: 12,306
Trainable params: 12,306
Non-trainable params: 0
    
```

Fig. 21. Structure of the Proposed DL Model.

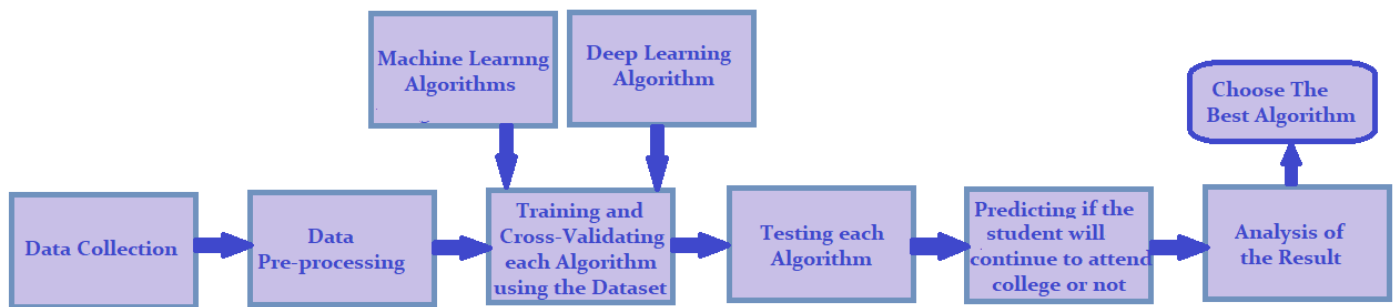


Fig. 22. Methodology for the Prediction if Student will Continue to Attend University or not.

IV. RESULTS AND DISCUSSION

A. Performance Evaluation

We used the most popular performance measures for machine and deep learning algorithms such as: precision, recall, accuracy, and F1-score as outlined in eq. 1, 2, 3, and 4.

$$\text{Accuracy} = (\text{TP} + \text{TN}) / (\text{TP} + \text{TN} + \text{FP} + \text{FN}) \quad (1)$$

$$\text{Precision} = (\text{TP} / (\text{TP} + \text{FP})) \quad (2)$$

$$\text{Recall} = (\text{TP} / (\text{TP} + \text{FN})) \quad (3)$$

$$\text{F1-score} = 2 \times (\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall}) \quad (4)$$

Where TP = True Positive, TN = True Negative

FP = False Positive, FN = False Negative

B. Performance of used Algorithms

We used a group of machine learning and one proposed deep learning algorithm for the prediction of whether a student will continue to attend university or not.

The machine learning algorithms belong to different categories of machine learning like Naïve based, SVM, KNN, trees, analysis, and others [28-30]. The proposed deep learning algorithm is custom made for the prediction if a student will continue to attend university or not.

We have split the dataset into Training, Validating, and Testing. We have trained every algorithm individually using our training dataset, tested it and we made a record of its

accuracy, recall, F1-score, precision, and time needed for training and testing.

Furthermore, we have trained the proposed deep learning algorithm using the same training dataset and cross-validated it using the validating dataset. We kept training the proposed DL algorithm until there was no room for improvement. We made a record of the DL algorithm accuracy, recall, F1-score, precision, and time needed for training, validating and testing. Part of the DL algorithm training, validating accuracies and losses are shown in Fig. 23.

It turned out that the best Machine Learning algorithm was NuSVC for predicting whether students will continue to attend university or not.

NuSVC (Nu-Support Vector Classification) belongs to the family of Support vector machines (SVM). SVM can be used in classification problems as well as regression problems. SVC (C-Support Vector Classification): Support vector classification, based on libsvm, the time complexity of data fitting is the second power of the data sample, which makes it difficult to expand to 10,000 dataset, when the input is multi-category (SVM was originally to deal with two classification problems), through a one-to-one solution, of course there are other solutions.

NuSVC core support vector classification, similar to SVC, is also implemented based on libsvm, but the difference is the number of support vectors through a parameter null value.

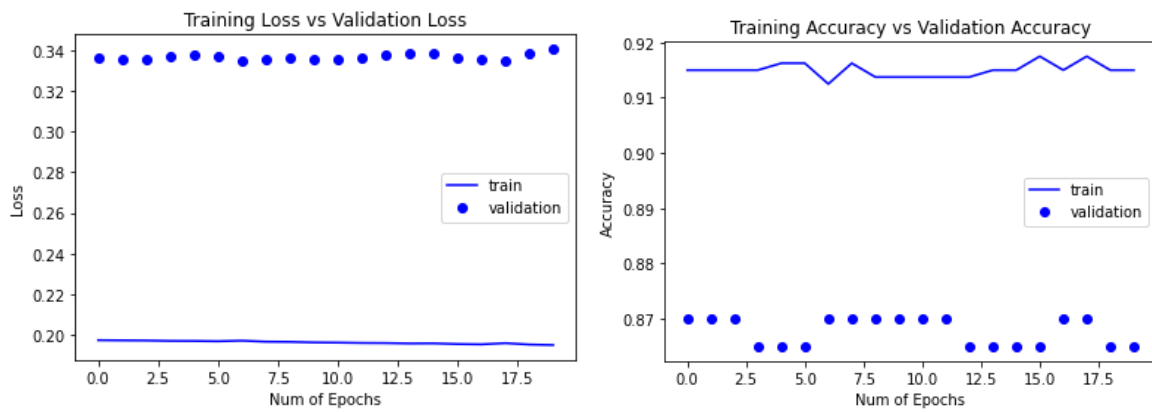


Fig. 23. Training vs Validation Losses and Accuracies.

NuSVC model achieved F1\_score (90.32%), Accuracy (91.00%), Recall (92.31%), Precision (88.42%), time required for training and testing (0.05 second) as can be seen in Table II.

Additionally, the proposed DL algorithm scored: Accuracy (93.23 percent), F1-score (93.05 percent), Recall (93.22 percent), Precision (92.55 percent), and time necessary for testing (0.67 second) for predicting whether student will continue to attend university or not as can be seen in Table III.

The first aim of this study was to answer the following question: Can we use the dataset at hand to create a supervised

machine learning classifier to predict reliably whether a student will go to university or not? The answer for this question is yes, we were able to create a supervised machine learning algorithm and F1-score accuracy was 90.32%.

The second aim of this study was: Can we use the dataset at hand to create a supervised Deep learning classifier to predict reliably whether a student will go to university or not?

The answer to this question was also yes and achieved an F1-score of 93.05%.

TABLE II. PERFORMANCE OF THE MACHINE LEARNING ALGORITHMS

Machine Learning Model-Name	Accuracy	Precision	Recall	F1_score	Time-in-Sec
NuSVC	91.00%	88.42%	92.31%	90.32%	0.05
RandomForestClassifier	90.00%	90.00%	90.00%	90.00%	0.19
GradientBoostingClassifier	89.50%	87.23%	90.11%	88.65%	0.14
LogisticRegressionCV	88.00%	83.84%	91.21%	87.37%	0.31
QuadraticDiscriminantAnalysis	87.50%	83.00%	91.21%	86.91%	0.01
MLPClassifier	87.50%	84.38%	89.01%	86.63%	0.68
LogisticRegression	87.00%	82.18%	91.21%	86.46%	0.01
LGBMClassifier	87.50%	85.87%	86.81%	86.34%	0.08
LinearSVC	86.50%	81.37%	91.21%	86.01%	0.01
CalibratedClassifierCV	86.50%	81.37%	91.21%	86.01%	0.06
LinearDiscriminantAnalysis	86.50%	82.65%	89.01%	85.71%	0.01
AdaBoostClassifier	86.00%	80.58%	91.21%	85.57%	0.10
GaussianNB	86.50%	84.04%	86.81%	85.41%	0.00
BaggingClassifier	86.50%	84.04%	86.81%	85.41%	0.05
Perceptron	85.00%	79.05%	91.21%	84.69%	0.01
SGDClassifier	84.50%	81.25%	85.71%	83.42%	0.01
KNeighborsClassifier	84.00%	84.71%	79.12%	81.82%	0.01
LabelPropagation	83.50%	82.22%	81.32%	81.77%	0.03
ExtraTreeClassifier	81.50%	78.12%	82.42%	80.21%	0.01
DecisionTreeClassifier	80.50%	74.53%	86.81%	80.20%	0.01

TABLE III. PERFORMANCE OF THE PROPOSED DEEP LEARNING ALGORITHM

Deep Learning Model-Name	Accuracy	Precision	Recall	F1_score	Time-in-Sec
Proposed DL Model	93.23%	92.55%	93.22%	93.05%	0.67

The last aim of the study was: Would the performance of Deep Learning techniques be better than the machine learning techniques in this case?

The answer is yes, because the deep learning algorithm scored 93.05% in F1-score while the machine learning algorithm achieved 90.32%.

#### V. LIMITATION OF THE STUDY

The current study is limited in terms of the dataset collected where it consists of 11 features (10 input features and one output feature) and 1100 records only. Furthermore, the current study is limited in terms of ML algorithms used, where we used only 20 ML algorithms among many as indicated in Table II.

#### VI. CONCLUSION AND FUTURE WORKS

In this study, we used twenty Machine Learning algorithms and a deep learning algorithm for predicting whether students will continue to attend university or not. The dataset was collected from Kaggle Repository. In order to predict whether students will continue to attend university or not, a group of 20 machine learning and one deep learning algorithm were used. Among the machine learning models used, the best machine-learning algorithm was NuSVC for predicting whether students will continue to attend university or not. NuSVC model achieved F1-score (90.32 percent), Accuracy (91.00 percent), Recall (92.31 percent), Precision (88.42 percent), time required for training and testing (0.05 second). Furthermore, the proposed DL algorithm attained: F1-score (93.05 percent), Accuracy (93.23 percent), Recall (93.22 percent), Precision (92.55 percent), time required for training and testing (0.67 seconds) for predicting whether student will continue to attend university or not.

In future work, other methods of ML algorithms may be utilized and the deep learning model should be tuned further to get better performance.

#### REFERENCES

- [1] Basem S. Abunasser, Mohammed Rasheed J. AL-Hiealy, Ihab S. Zaout and Samy S. Abu-Naser, "Breast Cancer Detection and Classification using Deep Learning Xception Algorithm" International Journal of Advanced Computer Science and Applications(IJACSA), vol. 13, no. 7, 2022.<http://dx.doi.org/10.14569/IJACSA.2022.0130729>.
- [2] Basem S. Abunasser, Mohammed Rasheed J. AL-Hiealy, Alaa M. Barhoom, Abdelbaset R. Almasri and Samy S. Abu-Naser, "Prediction of Instructor Performance using Machine and Deep Learning Techniques" International Journal of Advanced Computer Science and Applications(IJACSA), vol. 13, no. 7, 2022. <http://dx.doi.org/10.14569/IJACSA.2022.0130711>
- [3] Alkayyali, Z. K., et al. "Prediction of Student Adaptability Level in e-Learning using Machine and Deep Learning Techniques." International Journal of Academic and Applied Research (IJAAR), vol. 6, no. 5, pp. 84-96, 2022.
- [4] Obaid, T., Eneizan, B., Naser, S.S.A., ...Abualrejal, H.M.E., Gazem, N.A. Factors Contributing to an Effective E- Government Adoption in Palestine. Lecture Notes on Data Engineering and Communications Technologies, vol. 127, pp. 663–676, 2022.
- [5] Saleh, A., Sukaik, R., Abu-Naser, S.S. Brain tumor classification using deep learning. Proceedings - 2020 International Conference on Assistive and Rehabilitation Technologies, iCareTech 2020, pp. 131–136, 2020.
- [6] Alsaqqa, A. H., et al. "Using Deep Learning to Classify Different Types of Vitamin." International Journal of Academic Engineering Research (IJAER), vol. 6, no. 1, pp. 1-6, 2022.

- [7] Arqawi, S., Atieh, K.A.F.T., Shobaki, M.J.A.L., Abu-Naser, S.S., Abu Abdulla, A.A.M. Integration of the dimensions of computerized health information systems and their role in improving administrative performance in Al-Shifa medical complex, Journal of Theoretical and Applied Information Technology, vol. 98, no. 6, pp. 1087–1119, 2020.
- [8] Alkronz, E. S., et al. "Prediction of Whether Mushrooms are Edible or Poisonous Using Back-propagation Neural Network." International Journal of Academic and Applied Research (IJAAR), vol. 3, no. 2, pp. 1-8, 2019.
- [9] Mady, S.A., Arqawi, S.M., Al Shobaki, M.J., Abu-Naser, S.S. Lean manufacturing dimensions and its relationship in promoting the improvement of production processes in industrial companies. International Journal on Emerging Technologies, vol. 11, no. 3, pp. 881–896, 2020.
- [10] Albatish, I.M., Abu-Naser, S.S. Modeling and controlling smart traffic light systems using a rule based system. Proceedings - 2019 International Conference on Promising Electronic Technologies, ICPET 2019, pp. 55–60, 2019.
- [11] Elzamy, A., Messabia, N., Doheir, M., ...Al-Aqqad, M., Alazzam, M. Assessment risks for managing software planning processes in information technology systems. International Journal of Advanced Science and Technology, vol. 28, no. 1, pp. 327–338, 2019.
- [12] Abu Ghosh, M.M., Atallah, R.R., Abu Naser, S.S. Secure mobile cloud computing for sensitive data: Teacher services for palestinian higher education institutions. International Journal of Grid and Distributed Computing, vol. 9, no. 2, pp. 17–22, 2016.
- [13] Elzamy, A., Hussin, B., Naser, S.A., ...Selamat, A., Rashed, A. A new conceptual framework modeling for cloud computing risk management in banking organizations. International Journal of Grid and Distributed Computing, vol. 9, no. 9, pp. 137–154, 2016.
- [14] Alfarrar, A. H., et al. "Classification of Pineapple Using Deep Learning." International Journal of Academic Information Systems Research (IJAISR), vol. 5, no. 12, pp. 37-41, 2021.
- [15] Naser, S. S. A. JEE-Tutor: An intelligent tutoring system for java expressions evaluation. Information Technology Journal, vol. 7, no. 3, pp. 528-532, 2008.
- [16] Taha, A. M., et al. "Gender Prediction from Retinal Fundus Using Deep Learning." International Journal of Academic Information Systems Research (IJAISR), vol. 6, no. 5: pp. 57-63, 2022.
- [17] Naser, S. S. A. Developing an intelligent tutoring system for students learning to program in C++. Information Technology Journal, vol. 7, no. 7, pp. 1051-1060, 2008.
- [18] Salman, F. M., et al. "COVID-19 Detection using Artificial Intelligence." International Journal of Academic Engineering Research (IJAER) vol. 4, no. 3, pp. 18-25, 2020.
- [19] Naser, S. S. A. Developing visualization tools for teaching AI searching algorithms. Information Technology Journal, vol. 7, no. 2, pp. 350-355, 2008.
- [20] Abu-Jamie, T. N., et al. "Six Fruits Classification Using Deep Learning." International Journal of Academic Information Systems Research (IJAISR), vol. 6, no. 1, pp. 1-8, 2022.
- [21] Barhoom, A. M., et al. (2022). "Diagnosis of Pneumonia Using Deep Learning." International Journal of Academic Engineering Research (IJAER), vol. 6, no. 2, pp. 48-68, 2022.
- [22] Naser, S. S. A. Intelligent tutoring system for teaching database to sophomore students in Gaza and its effect on their performance. Information Technology Journal, vol. 5, no. 5, pp. 916-922, 2006.
- [23] Buhisi, N. I., & Abu-Naser, S. S. Dynamic programming as a tool of decision supporting. Journal of Applied Sciences Research, vol. 5, no. 6, pp. 671-676, 2009.
- [24] Ashqar, B. A. M., et al. "Identifying Images of Invasive Hydrangea Using Pre-Trained Deep Convolutional Neural Networks." International Journal of Academic Engineering Research (IJAER) vol. 3, no. 3, pp. 28-36, 2019.
- [25] Abu Naser, S.S. Evaluating the effectiveness of the CPP-Tutor, an intelligent tutoring system for students learning to program in C++. Journal of Applied Sciences Research, vol. 5, no. 1, pp. 109-114, 2009.



- [26] Fawzy Alsharif, Safi Safi, Tamer AbouFoul, Mustafa Abu Nasr, Samy Abu Nasser. Mechanical Reconfigurable Microstrip Antenna. International Journal of Microwave and Optical Technology, vol. 11, no. 3, pp.153-160, 2016.
- [27] Abu-Naser, S.S., El-Hissi H., Abu-Rass, M., & El-khozondar, N. (). An expert system for endocrine diagnosis and treatments using JESS. Journal of Artificial Intelligence, vol. 3, no. 4, pp. 239-251, 2010.
- [28] Aish, M. A., et al. "Classification of pepper Using Deep Learning." International Journal of Academic Engineering Research (IAER), vol. 6, no. 1, pp. 24-31, 2022.
- [29] Elzamly, Abdelrafe and Hussin, Burairah and Abu Naser, Samy and Doheir, Mohamed (2015) *Classification of Software Risks with Discriminant Analysis Techniques in Software planning Development Process*. International Journal of Advanced Science and Technology, 81 (2015). pp. 35-48. ISSN 2005-4238.
- [30] Abu-Saqer, M. M., et al. "Type of Grapefruit Classification Using Deep Learning." International Journal of Academic Information Systems Research (IAISR), vol. 4, no. 1, pp. 1-5, 2020.
- [31] Elzamly, Abdelrafe and Hussin, Burairah and Abu Naser, Samy and Doheir, Mohamed (2015) *Predicting Software Analysis Process Risks Using Linear Stepwise Discriminant Analysis: Statistical Methods*. International Journal of Advanced Information Science and Technology, 2015 (June). pp. 108-115. ISSN 2319-2682.
- [32] Almadhoun, H. et al. "Classification of Alzheimer's Disease Using Traditional Classifiers with Pre-Trained CNN." International Journal of Academic Health and Medical Research (IAHMR), vol. 5, no. 4, pp.17-21. 2021.

# Using the Agglomerative Hierarchical Clustering Method to Examine Human Factors in Indonesian Aviation Accidents

Based on the National Transportation Safety Committee (KNKT) Database 1997-2020

Rossi Passarella<sup>1\*</sup>, Gulfi Oktariani<sup>2</sup>, Dedy Kurniawan<sup>3</sup>, Purwita Sari<sup>4</sup>

Department of Computer Engineering, Faculty of Computer Science. Universitas Sriwijaya, Indralaya 30662, Indonesia<sup>1,2</sup>  
Intelligent System Research Group, Faculty of Computer Science, Universitas Sriwijaya, Palembang 30139, Indonesia<sup>1</sup>  
Informatic Management, Faculty of Computer Science. Universitas Sriwijaya, Indralaya 30662, Indonesia<sup>3,4</sup>

**Abstract**—This study aims to provide a comprehensive source of knowledge regarding aviation accidents in Indonesia caused by human factors, which are the most significant among other causative elements, requiring a detailed assessment of the accident as a result of pilot and co-pilot faults while operating the aircraft. The KNKT website database is still in the form of accident reports. To this end, the retrieved information based on historical data for 23 years of accidents caused by humans by analyzing the data using the clustering approach to gain data insight in the relationship between total flying hours and pilot licenses. The data analysis revealed that, in general, the aircraft operator complied with the CASR standards.

**Keywords**—Aviation accidents data; pilot's licenses; flying hours; human factor

## I. INTRODUCTION

The advancement of technology, particularly in the transportation industry, is usually swift; and one mode of transportation is nearly numbered one in aviation. Due to the relatively rapid time efficiency, security, and safety, aircraft make it easier for individuals to travel between provinces and across nations [1]. However, using this mode of transportation has been associated with a great risk of accidents [2][3]. Several studies on aviation risk have been conducted, especially those related to flight crews [4] [5].

Human casualties, environmental damage, property loss, and psychological impacts are all losses caused by plane crashes. According to Article 3 paragraph (1) of The Minister of Transportation (Indonesia) Regulation No. 77 of 2011, compensation for passengers who perished in aviation accidents amounted to Rp. 1,250,000,000 (one billion two hundred and fifty million rupiahs) per passenger, which is paid to the heirs of the deceased [6]. According to study [3] [7], four elements are found responsible for plane accidents: human factors, environmental factors, facility issues, and engineering considerations. Human factors affecting pilots and co-pilots, aviation security officials, and poor aircraft maintenance staff are among the four. This is followed by environmental conditions such as dense clouds, heavy rain, high winds, and mountains. The properties of the runway and the potential risk of animals being found on it contribute to the facility issues mentioned. Finally, engineering factors are associated with the

type of engine utilized as well as aircraft and engine maintenance.

According to study [7], statistics on aviation accidents and incidents in Indonesia gathered from the National Transportation Safety Committee (KNKT) show 26 significant incidents and 15 accidents between 2010 and 2016. There has been a 20% distribution of all events in the last seven years, and several factors were found responsible for such events. The percentage calculation of various causal elements, such as human factors (67.12 percent), technical factors (15.75 percent), environmental factors (12.33 percent), and facilities (4.79 percent) was obtained from the research. Given this context, investigators looked into whether human factors were still relevant after 2016, where to analyze data from the same sources between 2017 and 2020, which revealed that human factors were still the dominant factor in accidents and incidents, accounting for 52.6 percent of all accidents and incidents. This percentage represents a decrease of approximately 14.52 percent from the previous data.

Furthermore, a preliminary experiment was conducted using the same data provided by the KNKT (1997-2020), and using the same approach it was found that the human component still played a role in accidents 23 years ago with a percentage of 46.5 percent [7] [8]. Consequently, if the pilot's performance does not meet the criteria, does not correspond to a standard operating procedure (SOP), or does not correspond to the flying hours, the passenger will be at risk. As a result, the requirements to become a pilot must be understood, of which their license is the most important.

According to [8], the terms of the pilot's license are based on total flight hours under the provisions of the Civil Aviation Safety Regulation (CASR). There are four main licenses in aviation. For a prospective pilot to fly while they are learning, one must receive a Student Pilot License (SPL). When the pilot's flight hours increase, the pilot gets a Private Pilot License (PPL) with a minimum standard flight duration of 50 hours. To earn a Commercial Pilot License (CPL), a minimum standard flight duration of 150 hours must be met, and to obtain an Airline Transport Pilot License (ATPL), a minimum total flight duration requirement of 1500 hours must be met.

\*Corresponding Author.

An airline must determine the pilot in charge of flying a commercial aircraft far before the day of departure. What concerns us here is whether accident data from 1997 to 2020 contains improper pilot assignments due to licenses that were inappropriate for flying commercial aircraft flown, which may have led to the accidents. This issue is investigated to uncover the same. To this end, the retrieved information based on historical data for 23 years of accidents caused by human factors by analyzing the KNKT data using the clustering approach to gain data insight based on the variables' total flying hours and pilot licenses. Based on the description above, it is important to categorize total flight hours and pilot licenses such that they may be considered objects with characteristics to analyze.

A structural and hierarchical approach is required to address the research question, categorizing data based on pilot licenses who have encountered accidents or incidents. The agglomerative hierarchical clustering (AHC) method helps generate output in the form of a hierarchical structure that can provide an overview of data comprehensively. AHC is a bottom-up method wherein each data point is initially its own cluster, and as one ascends the hierarchy, additional clusters are amalgamated. In this method, the two neighboring clusters are merged into a single cluster. With this advantage, it allows for the researcher to identify hidden data from clusters so that they may view the distribution of the data and identify whether there are pilots in the KNKT data who have total flight hours.

This research only focuses on archival data of official aircraft accident reports issued by the KNKT Institute and finds out whether the human factor (pilot) shows a correlation with the cause of the accident. While factors outside the discussion are the limitations of this paper.

The structure of the presentation of this research is as follows: Section II describes the materials and methods used; Section III presents the results and discussion; and the conclusions are presented in Section IV.

## II. MATERIALS AND METHOD

### A. Materials

Researchers used credible and public data sources to obtain information on aviation accidents in Indonesia via the KNKT website database, which is still in the form of accident reports. This collection of accident reports is available on the website, organized by the year of the accident. The report data for each accident includes information on the event's timeline, the pilot's identity, and detailed information regarding the aircraft involved. However, as the data have not been integrated into a unified tabular format, researchers must consolidate accident data into a single database to simplify the analysis.

The researchers only used 341 data points from Indonesian aircraft accidents that occurred between December 19, 1997, and September 15, 2020, accounting for the completeness of each data point. The collected data is then analyzed to produce cluster results and conclusions regarding the link between two factors in an aircraft disaster.

### B. Method

The methodology carried out in this research process is described within a research framework, which is illustrated in Fig. 1.

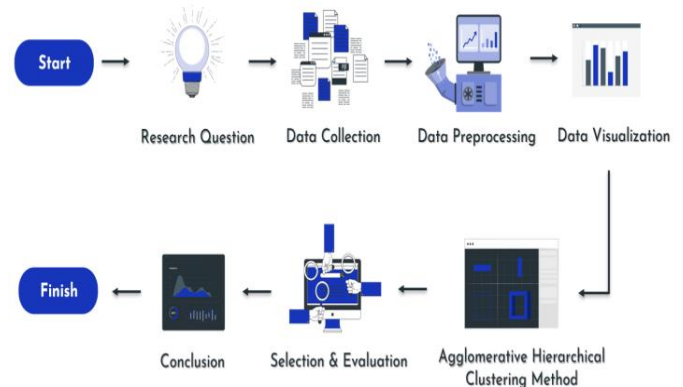


Fig. 1. Research Framework.

The first step of conducting any work of research is to develop a research topic. In making this research inquiry, the goal is to construct a question that is clear, focused, and free of prejudice. This is followed by providing a literature review to comprehend the context of the topic sufficiently. A literature review allows the researcher to construct questions that express the concepts and problems to be explored, also informing the public of the relevance of the outlined issues, which are beneficial to society once addressed, furthering scientific development.

In the second phase of the research, the data used in this study is explained in the previous sub-chapter; data from KNKT about aviation accidents in Indonesia have been used. Data collection of accidents is carried out and then put together into a spreadsheet. Researchers collect essential and relevant information and arrange them in the chronology of accidents. The data collected from accident reports have 16 variables, which include the date, year, province, location, aircraft type, aircraft operator, crew, passengers (PAX), injured, fatalities, type of accident (A/I), flight phase (POA), cause of the accident (PCOA), total pilot's flight hours, type of pilot's license and remarks, altogether culminating in 341 lines of data. After data collection, information is saved with the file extension .csv and imported into the Jupyter notebook program for data analysis.

The third stage is data preparation, which includes multiple processes such as data cleaning, data integration, data transformation, and data reduction. This step involves loading the data into the Jupyter notebook. Data processing is performed, followed by a data cleaning process, which includes the selection of the required variables. Following identification of the required variables, data that includes null or empty values will be cleaned up and outliers in the data will be verified and removed in case they impact the data balance.

The next stage in this study is data transformation. The encoder label is applied in the pilot license variable column to change the data from category to numeric. Furthermore, data reduction will complete the standard scaler procedure at this step, which is required to normalize the value of the variable column of total flight hours and pilot license. Data will be visualized using EDA after completing the data pretreatment stage as part of the pre-modeling phase to display variable data such as total flying hours and pilot licenses, following which, the clustering method is employed for analysis.

The researchers adopted the AHC approach at this stage. This clustering method is divided into several sub-methods for the calculation of proximity between clusters. Researchers only use three methods, which are single-linkage (Closest distance), complete linkage (Farthest distance), and average linkage (Average distance). It then uses Euclidean calculations to compute distance and the silhouette index to estimate the ideal number of clusters. Two variables are employed in this grouping: total flight hours and pilot's license, which will be clustered to determine the link between the two and the frequency of accidents that occur as a result.

Following the above mentioned procedure, a validation test on the number of each cluster found is required to select the best approach. This selection will be based on the value of the validation test index that has a satisfactory performance by analyzing the distribution of the data for each cluster. This is followed by an analysis of research questions based on the relationship between the two variables used, namely total flight hours and pilot license. The method is then evaluated for selection.

The final stage is the conclusion, which will include the results of the AHC method, which involves a validation test value that meets the criteria with the number of clusters used. The results are bound to demonstrate whether or not a relationship relevant to the research question can be established. Furthermore, significant insights will be discovered in the form of information due to the preceding stage's clustering and data distribution.

### III. RESULT AND DISCUSSION

#### A. Data Preparation

Researchers execute variable filters on this preparation data used in this investigation. It is known that there are 16 variables from raw data; however, in this study procedure, only two variables are employed, while the remaining variables are not used. Both variables, total flight hours and pilot's license are utilized.

#### B. Data Preprocessing

The data on aviation accidents in Indonesia from 1997 to 2020 were collected using the date of the accident reports on KNKT websites. The first step is to erase the missing data values in the variable column used, particularly the total flight hours and pilot's license. After deletion, the total data collected was 177 out of 341 reports. The second procedure is to remove outliers - data with distinctive features that appear different from other data. After removing missing data values, outliers acquired 156 data points from a total of 177 data points.

TABLE I. DATA BASED ON PILOT'S LICENSE

No.	Pilot Licenses	Numbers of Data
1	ATPL	99
2	CPL	45
3	SPL	8
4	PPL	4

Data that has already been cleaned will be transformed into a suitable form for processing. The first step is to examine the various types of pilot licenses including ATPL, CPL, SPL, and PPL. The results are shown in Table I, with data from four pilot licenses.

This step is followed by the creation of an encoder label for the type of pilot license. ATPL (0), CPL (1), PPL (2), and SPL (3) are the designations for each type of license. In this study, data reduction is used to reduce dimensions in data so that later on, data will be comprehensive; this would also retain the integrity of the data as much as possible. This reduction data strategy is a standard scaler that seeks to normalize the data so that massive variations may be avoided. It is used in this study to standardize the variables - total flight hours and pilot's license value.

#### C. Descriptive Statistics

Following the completion of the data preparation stage, descriptive statistics of the data were used as an additional tool within the data processing stage before proceeding to the data visualization stage. This study employs two types of variables: quantitative and qualitative data variables. The quantitative data variables in this study are variables with numerical numbers - the total flight hours of pilots. Qualitative data variables are categorical variables - pilot licenses. The results are produced using descriptive statistics based on the quantitative and qualitative data factors, as shown in Table II.

According to Table II, the mean or average value of the accurate data for the quantitative data variable (total pilot flight hours) acquired is 7,024.58 hours. Furthermore, for the dispersion value (data spread to the average value) acquired by 0.69, this variable's median value was found to be 5,935 hours. This signifies that the variable corresponds with homogenous data with a tiny dispersion value - a minimum value of 20 hours and a maximum value of 17,547 hours. It also implies that the missing value for this variable is already worth 0, which means that there is no lost or empty data. While the type of qualitative data variable (pilot license) is categorical, it is converted to numerical using the encoder label. The pilot license data has a mean value of 0.50.

Furthermore, the median value of this variable is 0, indicating that ATPL is the midpoint value of the pilot license variable. Aside from the dispersion value, the average value was found to be 1.59, indicating that this variable corresponds to heterogeneous data with a high dispersion value, a minimum value of 0, and a maximum value of 3. For this variable, a missing value is already worth 0, indicating that no data was lost or is empty.

TABLE II. DESCRIPTIVE STATISTICS OF DATA

Variable Types	Mean	Median	Dispersion	Min	Max	Missing Value
Quantitative	7024.58	5935	0.69	20	17547	0(0%)
Qualitative	0.50	0	1.59	0	3	0(0%)

D. Data Visualization (EDA)

Following the data preprocessing and descriptive statistics phases, the EDA step is used to comprehend the contents and components of the data. The EDA approach employed in this study is univariate and bivariate; univariate analysis is used to distribute data from a single variable while bivariate analysis demonstrates the link between two variables [9]. To comprehend data trends, both analyses are used to examine the distribution of data for varying total flight hours and pilot licenses in detail. As follows, pie charts are used for univariate analysis and scatter plots are used for bivariate analysis.

The Pie Chart is used to determine how much category distributes on a pilot's license. According to Fig. 2, the most significant proportion of ATPL (0) license types was 63.2 percent, followed by 29.0 percent for the CPL (1) category, 2.6 percent for the PPL (2) category, and 5.2 percent for the SPL (3) type. The scatter plot is used to illustrate data distribution, to examine how data spread links between the variable total flight hours and the pilot's license.

According to Fig. 3, it may be noted that the patterns across the two variables do not have a link as no tendency of specific values can be observed in common. Each pilot's license has its unique pattern that appears to be completely unrelated to the value of each pilot's license in the crash data.

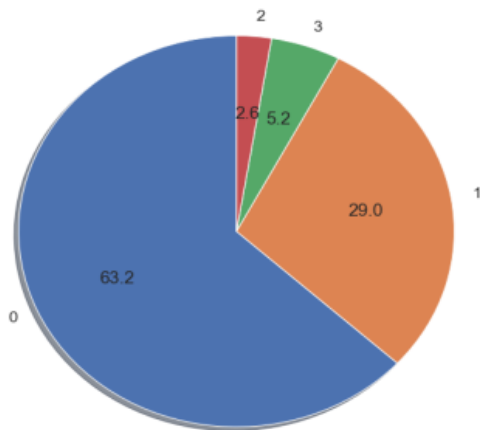


Fig. 2. Pilot License Data Visualization.

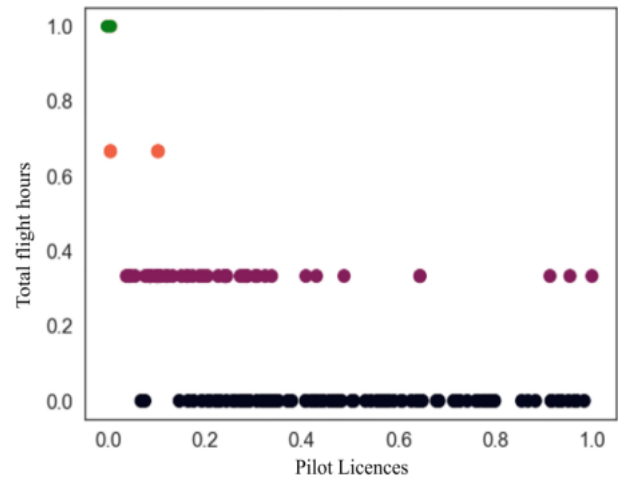


Fig. 3. Visualization of Total Flight Hours based on Pilot's License.

E. Cluster Test

At this juncture, it is important to mention that the test used three predefined agglomerative strategies. First, the distance is calculated using the Euclidean formula, then the ideal number of clusters is estimated using the silhouette index technique, the amount of data spread for each cluster is examined, and finally, the dendrogram of each approach is visualized.

The following are the findings of the validation test of the number of clusters using the silhouette index technique with a range of cluster values 2-4. Using two clusters yielded an average value of 0.5156, three clusters yielded an average value of 0.5004, and four clusters yielded an average value of 0.5500. Based on the computation of the silhouette index, if the average value is close to the value of one, the data grouping is considered better [10].

F. Hierarchical Agglomerative Method Test

The grouping approach employs the three techniques selected to acquire the data dispersed across clusters and provide dendrogram visualizations based on the number of clusters formed from each AHC method.

1) *Single linkage*: The single linkage approach is used to group two variables with a distance between cluster pairs determined by the two nearest objects between separate clusters [11]. According to Table III, each cluster contains a varied quantity of data, and the data in each cluster is far apart and unbalanced. Cluster 3 has the least data with 4 data points, while Cluster 2 has the most data with 99 data points. The dendrogram construction employs a single linking approach with four clusters connected by a pair of points with the closest distance between each other, as depicted in Fig. 4. (a).

TABLE III. NUMBER OF EACH SINGLE LINKAGE CLUSTER

Cluster	Numbers of Data
1	45
2	99
3	4
4	8

2) *Complete linkage*: The distance between clusters has been computed on the furthest distance between a pair of items [12] [13]; the quantity of data per cluster and visualization results (using a dendrogram) was determined with the Euclidean formulae for four clusters. The data of each cluster corresponded with a ratio that was less balanced than the single linkage technique due to the quantity of data used in the complete linkage method (See Table IV). Cluster 4 has the least quantity of data (12) while Cluster 1 has the most data (72).

TABLE IV. NUMBER OF EACH COMPLETE LINKAGE CLUSTER

Cluster	Numbers of Data
1	72
2	42
3	30
4	12

The output of the dendrogram implementation employs the entire linkage approach, with four clusters connected from the longest distance from a pair of objects seen in Fig. 4. (b).

3) *Average linkage*: The average linkage computes the average distance between all pairs of data points from various clusters [14]. The amount of data per cluster and the visualization results used for four clusters were determined using Euclidean equations. The number of clusters in this technique is not found to be highly significant, as seen in Table V. In comparison to the whole linkage and single linkage approaches, the amount of data in each cluster is more evenly distributed. Cluster 1 has the smallest amount of data, whereas Cluster 2 has the largest. The average linkage approach is used to achieve the results of the dendrogram implementation, as shown in Fig.4(c).

TABLE V. NUMBER OF EACH AVERAGE LINKAGE CLUSTER

Cluster	Numbers of Data
1	30
2	53
3	42
4	31

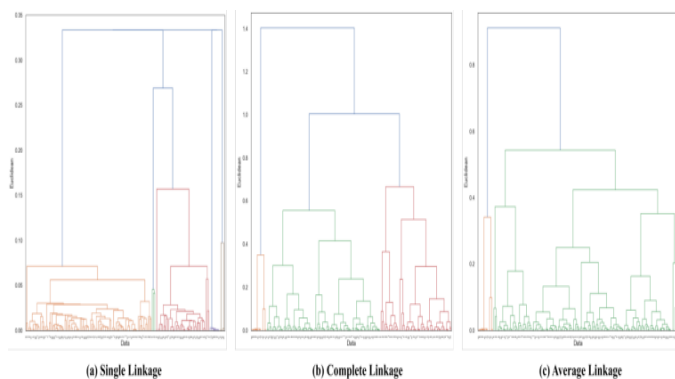


Fig. 4. Visualization Dendrogram Method: (a) Single-linkage (b) Complete Linkage (c) and Average Linkage.

### G. Index Validation Test

The validation test of the score value in clustering is used to determine the optimal AHC approach. The silhouette index matrix was used in the first measuring procedure, the Davies Bouldin index in the second, and the Calinski Harabasz index in the third.

The data grouping is considered better in silhouette index calculations if the index value is close to one [15]. The results of calculations using the silhouette index based on Table VI show that the index values for the three grouping techniques differ. Compared to the single and complete linkage, the average linkage technique yields the highest index value with 0.5623 or 56% being calculated. The silhouette index method was then used to obtain the best method on the validation test, namely average linkage.

The Davies Bouldin index is used for the second measurement. According to [16], the index value approaches zero, implying that the data grouping in one cluster is better. The Davies Bouldin index calculations in Table VI show that the three techniques have varying index values. The average linkage technique has the lowest index value of 0.5644 or 56%. As a result of the validation test using the Davies Bouldin index, the best technique obtained is average linkage.

The Calinski Harabasz index is used in the third measurement, wherein the index value is not limited. The higher the index value, the better the data grouping in one cluster [17]. Table VI shows the results of the Calinski Harabasz index calculation. It is well known that the index values of the three techniques differ. The average linkage technique produces the highest index value of 284.68, while the single linkage technique produces the lowest index value of 93.37. Using the Calinski Harabasz index, the best technique is determined using the validation test, which was found to be average linkage. For the three validation test methods, all the measurements showed elicited 100 percent, indicating that average linkage is the best method for viewing data insights into licenses and total flight hours.

TABLE VI. INDEX VALIDATION TEST RESULTS

Methods	Single	Complete	Average
Silhouette	0.4838	0.4045	0.5623
Davies Bouldin	0.6238	0.7421	0.5644
Calinski Harabasz	93.37	159.09	284.68

### H. Analysis of Cluster Average Linkage Results

This study shows the results of each cluster in relation to the total flight hours and pilot licenses based on the type of the pilot license used - ATPL, CPL, SPL, and PPL, which were mapped based on the number of clusters. The average value has also been based on the number of clusters. Table VII displays the number of pilot license categories in each cluster and the total number of pilots' flight hours.

After selecting the agglomerative method approach with average linkage, the researcher noticed two accidents in the fourth cluster involving pilot licenses under the ATPL category (Table VII). While the pilot involved in the fourth cluster had



an average flight time of 1071.67 hours, which when combined should have been more than 1071.67 hours, the ATPL standard is at least 1500 hours. Such anomalies must be investigated as the total rata-average of these four licensing classes is 1071.67 hours, while certain anomalies may be assumed to have a duration of over 1071.67 hours. Thus, a deeper delve into Cluster 4 was necessary. Based on the established statistics, researchers found two ATPL-type permits with a total flight hour of fewer than 1500 hours. Despite CASR standards requiring ATPL type pilots to have a minimum flight hour requirement of 1500 hours, two pilots in Cluster 4 did not meet the minimum flying hours required by their pilot's license, with total flight hours of 1200 hours and 1339 hours, for each pilot. The Cluster 4 results are shown in Fig. 5, illustrating the ATPL (0) licenses that were found to be different from the rest of the data.

TABLE VII. NUMBER OF PILOT'S LICENSES AND AVERAGE TOTAL PILOT HOURS

Cluster	Pilot Licenses				Average of Total Data Pilot's Flight (Hours)
	ATPL (0)	CPL (1)	PPL (2)	SPL (3)	
1	27	3	0	0	14625,23
2	33	20	0	0	4634,84
3	37	5	0	0	9060,54
4	2	17	4	8	1071,67

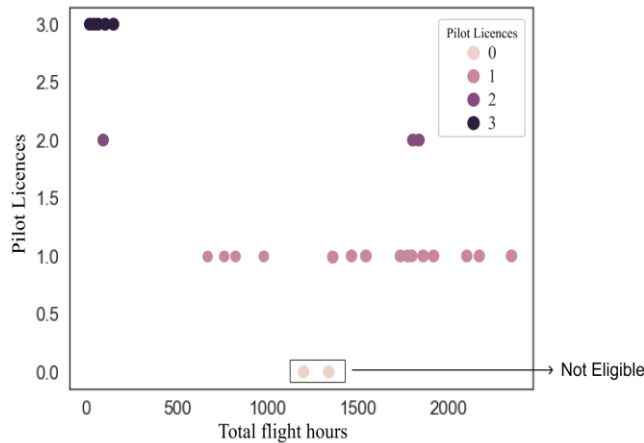


Fig. 5. Visualized Data Cluster 4 Average Linkage.

#### IV. CONCLUSION

This study was carried out to provide a comprehensive source of knowledge regarding aviation accidents in Indonesia caused by human factors, which are the most significant among other causative elements, requiring a detailed assessment of the accident as a result of pilot and co-pilot faults while operating the aircraft. The approach used to analyze KNKT data was clustering algorithms to identify insightful data in the relationship between total flying hours and pilot's license, based on the historical data of accidents caused by human factors over 23 years.

The data analysis revealed that, in general, the aircraft operator complied with the CASR standards. However, using clustering, it was discovered that the pilot had a license that did not match with the flown aircraft in 1.3 percent of the 156 accidents. Another finding revealed that, in comparison with other causes, human-caused accidents (pilots) had the highest proportion - with 46.5 percent from 1997 to 2020. Meanwhile, the level of representation of the analyzed data was only 46 percent as only 156 out of 341 accidents over 23 years were processed due to the relevant observation values on license data variables and flight hours.

#### ACKNOWLEDGMENT

We thank those who have helped us complete this study.

#### REFERENCES

- [1] R. Passarella, and S. Nurmaini, 2022 "Behavioral Evidence of Public Aircraft with Historical Data: The Case of Boeing 737 MAX 8 PK-LQP," *Journal of Applied Engineering Science*, vol.20, no.4.
- [2] Y. Wei, H. Xu, Y. Xue, and X. Duan 2020 "Quantitative assessment and visualization of flight risk induced by coupled multi-factor under icing conditions," *Chinese J. Aeronaut.*, vol. 33, no. 8, pp. 2146–2161. DOI: 10.1016/j.cja.2020.03.025.
- [3] R. Passarella, and S. Nurmaini 2022 "Data Analysis Investigation: Papua is The Most Unsafe Province in Indonesia for Aviation: An Exploratory Data Analysis Study from KNKT-Database Accidents and Incidents (1988-2021)," *Journal of Engineering Science and Technology Review (JESTR)*, vol. 15, no.3, pp 158-164. DOI: 10.25103/jestr.153.17.
- [4] S. Gentile, A. Furia, and F. Strollo, 2020 "Aircraft pilot licence and diabetes," *Diabetes Research and Clinical Practice*, vol 161, DOI: 10.1016/j.diabres.2020.108047.
- [5] M. Efthymiou, S. Whiston, JF. O'Connell, and GD. Brown, 2021 "Flight crew evaluation of the flight time limitations regulation," *Case Studies on Transport Policy*, vol 9, no. 1, pp 280-290, DOI: 10.1016/j.cstp.2021.01.002.
- [6] R. Kristiawan, Rolan, Abdullah 2018, "Faktor penyebab terjadinya kecelakaan kerja pada area penambangan batu kapur unit alat berat pt. semen padang," *J. Bina Tambang*, vol. 5, no. 2, pp. 11–21.
- [7] Rahimudin 2015, "Analisis Faktor-Faktor Penyebab Kecelakaan Pesawat Udara Komersil Di Indonesia Pada Tahun 2002 Sampai Dengan Tahun 2012," vol. 8, pp. 82–83.
- [8] Y. Xue and G. Fu 2018, "A modified accident analysis and investigation model for the general aviation industry: emphasizing on human and organizational factors," *J. Safety Res.*, vol. 67, pp. 1–15. DOI: 10.1016/j.jsr.2018.09.008.
- [9] D. C.Hoaglin 2015, "Exploratory Data Analysis: Univariate Methods," *Int. Encycl. Soc. Behav. Sci. (Second Ed.)*, pp. 600–604. [Online]. Available: DOI:10.1016/B978-0-08-097086-8.42125-2.
- [10] K. Gonzalez, and S. Misra, 2022 "Unsupervised learning monitors the carbon-dioxide plume in the subsurface carbon storage reservoir," *Expert Systems with Applications*, vol. 201. DOI:10.1016/j.eswa.2022.117216.
- [11] F. Ros, and S. Guillaume, 2019 "A hierarchical clustering algorithm and an improvement of the single linkage criterion to deal with noise," *Expert Systems with Applications*, vol. 128, pp 96-108. DOI:10.1016/j.eswa.2019.03.031.
- [12] D. Krznaric, and C. Levcopoulos, 2002 "Optimal algorithms for complete linkage clustering in d dimensions," *Theoretical Computer Science*, vol. 286, no.1, pp 139-149. DOI:10.1016/S0304-3975(01)00239-0.
- [13] P. Govender and V. Sivakumar 2020, Application of k-means and hierarchical clustering techniques for analysis of air pollution: A review (1980–2019), vol. 11, no. 1. Turkish National Committee for Air Pollution Research and Control.

- [14] L. Ramos Emmendorfer and A. M. de Paula Canuto 2021, "A generalized average linkage criterion for Hierarchical Agglomerative Clustering," *Appl. Soft Comput.*, vol. 100, p. 106990. DOI: 10.1016/j.asoc.2020.106990.
- [15] M. Gagolewski, M. Bartoszek and A. Cena, 2021 "Are Cluster validity measure (in)valid ?", *Information Sciences*, Vol. 581,pp 620-636.
- [16] DL. Davies and DW. Bouldin, 1979 "A Cluster Separation Measure", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-1, No.2.
- [17] J. Baarsch and M.E. Celebi, 2012 " Investigation of Internal Validity Measures for K-Means Clustering". *Proceedings of the International MultiConference of Engineers and Computer Scientists (IMECS)*, Vol.1. March 14-16, Hong Kong.

# A Framework for Crime Detection and Diminution in Digital Forensics (CD3F)

Arpita Singh<sup>1</sup>, Sanjay K. Singh<sup>2</sup>  
Amity Institute of Information Technology  
Amity University  
Lucknow, India

Hari Kiran Vege<sup>3</sup>, Nilu Singh<sup>4</sup>  
Department of Computer Science and Engineering  
Koneru Lakshmaiah Education Foundation  
AP, India, Lucknow, India

**Abstract**—Cyber-attacks have become one of the world's most serious issues. Every day, they wreak serious financial harm to governments and people. As cyber-attacks become more common, so does cyber-crime. Identifying cyber-crime perpetrators and understanding attack tactics are critical in the battle against crime and criminals. Cyber-attack detection and prevention are difficult undertakings. Researchers have lately developed security models and made forecasts using artificial intelligence technologies to solve these concerns. In the literature, the authors explained numerous ways of predicting crime. They, on the other hand, have a problem forecasting cyber-crime and cyber-attack strategies. Here, in this paper author proposed a digital forensic investigation procedure that deals with cyber-crime. In this investigation, the process author explains digital forensics techniques for ensuring that digital evidence is located, collected, preserved, evaluated, and reported in such a way that the evidence's integrity is preserved. These sequential digital forensic stages affect a standard and accepted digital forensic investigation procedure, and each phase is influenced by sequential occurrences, with each event relying on tasks. Digital forensics investigation is a technique for ensuring that digital evidence is handled in such a way that the evidence's integrity is preserved. Sequential digital forensic stages affect a standard and accepted digital forensic investigation procedure, and each phase is influenced by sequential occurrences, with each event relying on tasks.

**Keywords**—Cyber-crime; digital forensics; digital evidence; data analysis; security and privacy; cyber-attack

## I. INTRODUCTION

Information security has given users complete control over data by specifying who has access to it, who can govern it, and who may receive it [1]. People's lifestyles are changing as a result of technological advancements. For example, nowadays most people prefer online payment to traditional payment, access to social media, medical consultation through phone or video chat, online schooling, and so on. As technology advances and new services become available, the number of internet users grows, and an exponential increase in information causes its use, as well as misuse, resulting in cyber-attacks and cyber-crime. Cyber-attacks have an impact on the economic systems of our countries. According to research by Mahindra SSG and ASSOCHAM, cybercrime costs India around 24,630 crores per year [2]. Attacks have become more complex as a result of technological advancements, and defending oneself is no longer as simple as installing anti-virus software.

The general concept of evidence preservation in the chain of custody has remained the same but the original process of investigation may vary. The preparation, examination, identification, collection, analysis, validation, acquisition, documentation, and forensic reporting of digital evidence in a court of law, is known as the digital forensics investigation process. The process of interpreting and imaging digital evidence from various electronic devices using scientifically sound and validated methodologies is an indestructible part of digital forensic investigation. Digital forensics is a rapidly growing field that uses a variety of analysis tools and computer investigative approaches to locate relevant legal evidence and hints [3]. In general, digital forensics is a process that involves not only retrieving information about a reported incident but also properly processing that information so that experts can obtain all relevant clues and evidence, which can then be used to pursue your legal interests against someone or in any other situation. Finding evidence, keeping it, accurately documenting it, and presenting it in a court of law are all part of the digital forensic process. Although, because this is not a simple process, it might often take years to solve the issue. Furthermore, complex systems are making it harder to hit these days. We now have a complex methodology and enhanced technologies and techniques to determine if any pieces of evidence are present or not. We now have a complicated methodology, as well as new technologies and techniques, to determine whether a criminal case has occurred and a huge amount of money is spent to solve the case [2].

Cyber-attacks and cyber-crimes are a serious worry for large countries such as the US and the UK, which have developed a number of security solutions to combat them [4]. All countries are seeking to secure and adapt to cyberspace security [5]. The security of critical infrastructure must be a top priority for countries [6]. In the year 2020, information taken from the Airbus Company's information system was sold on the dark web. Millions of people's medical information has been stolen, and several communities have declared a state of emergency as a result [7]. With each passing day, the workforce grows insufficient in combating cyber-attacks, necessitating the search for new alternatives. Machine-learning approaches are being used by researchers to detect power outages caused by cyber-attacks [8] and to prevent the Internet of Things vulnerabilities [9]. Other applications include detecting spam and network attacks [10], detecting phishing attempts against banks [11], and increasing sexual crimes on social media [12] Stock prediction [13], risk mapping [14], and

cyber profiling [15] are some of the sectors where these technologies have been used. Implementation areas include predicting crime trends and patterns [16], criminal identity detection [17], and crime prevention [18].

The authors suggest a framework called "Crime Detection and Diminution in Digital Forensics (CD3F)" in this study. The goal of this research article is to create a framework that connects the stable and sequential aspects of the digital forensic investigation process. With the many operations of the investigation workflow comprising physical and investigative duties and judgments. The Digital forensic framework's intended role is to enable effective, focused, and fast risk identification and management. This CD3F framework contains and outlines eight major workflow stages, as well as the procedures and duties that each stage entails. The article is organized as follows: some of the current digital forensic investigation process models proposed by the researchers are elaborated on and discussed, followed by an explanation of why the digital forensic workflow should be mapped. The following section is about "Phases involved in proposed framework and explanation" and gives an overview of the framework, including all eight stages that the author proposes. "The suggested digital forensic guiding framework specifics" is the topic of the study's next session which focused on the CD3F framework workflow stages: forensic request, preparation, examination, identification, collection, analysis, acquisition, and forensic reporting. The discussion and critical evaluation of the proposed framework are covered in greater depth in the paper. The text is then concluded with some notes about the study's significance and the last half of this research article.

## II. LITERATURE REVIEW

A description of all previously developed cyber forensics investigation frameworks that the authors investigated before developing the proposed framework is not possible due to space limits. Although some of the reviewed frameworks are described in this work, it should not be considered that the suggested model is based on them. In 2001 [19] authors proposed a framework where they have covered preparation, identification, permission, and communication were the four phases of the initial introduction model DFRWS. This paradigm was expanded into the SRDFIM framework, which includes additional steps such as scene securing, screening, scene documenting, evidence gathering, communication shielding, screening, preservation, analysis, and presentation [20]. The FaaS Framework 2014 is based on the IDFPM framework [21], which begins with the collecting and authentication stages. Evidence acquired throughout the investigation will be stored in central storage, followed by the inspection phase in this suggested architecture. The analysis phase will be conducted with current analysis tools, with the results being kept in a centralized database.

The DFRWS Investigation Model is used by the FBI. A fog IoT forensic framework (FOBI) is a network model that performs important operations such as data filtering and aggregation [22, 23]. Storage and processing resources are located at the network edge in this paradigm [24]. Fog protects data sent to IoT devices while also filtering traffic data. As a

result, this architecture provides a number of benefits to IoT devices, including reduced network latency, faster and smarter responsiveness, more scalability, and greater security and privacy. Early detection of a cyber-threat or cyber-attack. It can be used to discover problematic IoT devices by including a fog layer in the framework [22] [25]. Frameworks for research must also be adjusted as a result of technological advancements. As the author mentioned above, some frameworks are presented to combat crime using the most up-to-date technologies and techniques. With some rising issues, law agencies is required for a framework that can combat crime and track down criminals' tracks. Cloud computing is a new level of networking since it offers limitless processing capacity and storage, which poses security concerns [20]. Cyber-attacks such as distributed denial of service (DDoS) attacks that deliver harmful packets [26] and phishing attempts that trick users on banking and shopping sites have increased dramatically. Furthermore, attackers are increasingly deploying malicious attack software (viruses, worms, trojans, spyware, and ransomware) that is installed on a user's computer without their knowledge or agreement [28]. Social engineering attacks are, once again, the most widespread of these attacks and one of the most difficult to counter. They are based on technical expertise, ingenuity, and persuasion, and are carried out by exploiting the victim's weakness. Kevin Mitnick, a well-known hacker who specializes in social engineering attacks, was able to break into most of the computers he targeted using this method [29].

This attack is mentioned as one of the main security vulnerabilities in the system by Breda, Barbosa, and Morais [30], regardless of how secure a technical system is. Similarly, assaults on IoT devices, which have expanded dramatically in recent years, have a significant impact on society. For security reasons, assaults and threats to the IoT structure should be understood [31]. As described in this paper, studies performed to analyze and combat cyber-attacks highlight the importance of crime prediction. Many jurisdictions' legal frameworks characterize the attacks listed above as banned criminal offenses. The task of combating crime and criminals is delegated to law enforcement agencies. Researchers provide numerous analysis and prediction approaches to the institutions undertaking the research. Many studies, for example, have used big data [32] and machine-learning [18] methods to analyze crimes. With artificial intelligence models, they have contributed to crime and crime-fighting institutions. Identifying the regions where crime can be perpetrated and the story behind it [33], predicting crime using spatial, temporal, and demographic data [34], and assessing crime using literacy, unemployment, and development index data [35] are just a few examples. A time series of crime data from San Francisco, Chicago, and Philadelphia was utilized to forecast crimes in the year's ahead. K-nearest neighbors (KNN) and Naive Bayes (NB) classification models performed worse than Decision Trees (DT) [36]. Using the KNN and DTs, a crime prediction was made with an accuracy of 39 to 44 percent [38]. The location, kind, date, time, latitude, and longitude of crimes committed in the United States were used as input. The results of crime predictions using KNN Classification, Logistic Regression (LR), DTs, Random Forest (RF), Support Vector Machine (SVM), and Bayesian approaches showed that the

KNN classification was the most accurate at 78.9% [37]. Thirty-nine distinct categories of crime statistics from San Francisco were used in the study. A model splitting crimes into two types, blue/white-collar crime and violent/non-violent crime, was built using Gradient Boosted Trees and SVMs. The categorization of blue-white collar offenses was done with great precision.

The study, however, did not produce significant results in terms of classifying violent and non-violent crimes [39]. The data was taken from a ten-year murder in Brazil. The RF approach was used to make 97 percent accurate predictions in order to examine the effect of non-Gaussian residuals and urban metrics on murders. Unemployment and ignorance were found to be significant factors in homicide in this study. The relevance of each factor in predicting the crime was also assessed [40]. Another study employed the type, timing, and location of crime data to predict crime in certain Indian regions. It was decided to employ the KNN prediction algorithm. Robbery, gambling, accidents, violence, murder, and kidnapping crimes were predicted using this strategy. It was shown to be more successful than a previous study of a similar nature [41]. Using crime data obtained from social media networks, big data and machine-learning frameworks were created. Volunteered Geographic Information, web, and mobile crime reporting applications were used to collect the information. The NB algorithm was used to generate crime predictions from the collected data. The goal of these forecasts is to pinpoint the site of potential crimes so that they can be avoided [42].

The demographic and geographic data from past years' events were utilized to forecast terrorist attacks in India. Using artificial intelligence algorithms, this model predicted terrorist occurrences with a high degree of accuracy [43]. The data used to analyze cyber-crime was publicly available information from social media platforms such as Facebook and Twitter. The F-measure value, which is the degree of accuracy and precision, was used to compare the algorithms. The RF algorithm was shown to be the best fit in the circumstance, with an accuracy of 80%. Threats were identified automatically using a model that analyses cyber-crime [44]. Through the screening program, real-time crime data from the internet news was employed. The classification algorithms employed were SVM, Multinomial NB, and RF. The data was divided into two categories: criminal and non-criminal. The most essential aspect is that it now includes news analysis [45]. Machine-learning algorithms were used to classify data from cyber-crime incidents in India. The program, which was 99 percent accurate in predicting crimes, cut down on time spent on analysis and manual reporting [46]. Kaggle was utilized to obtain a universally compared intrusion detection dataset.

On the bases of the literature review, the authors observed and analyzed that cyber-attacks and crimes are vital to investigate since they inflict significant harm to persons and governments. The studies contributed significantly to the literature and, in particular, to the criminal investigation units. General crimes, cyber-crimes, and attacks are commonly employed as a dataset in these studies. The real dataset based on personal qualities is looked at to a lesser extent, and a framework for digital forensic inquiry is proposed as a result.

Because of the importance of the fields investigated, the cyber-attack and perpetrator estimating approach is addressed.

### III. PHASES INVOLVED IN PROPOSED FRAMEWORK AND EXPLANATION

Almost all major transactions are adopted by web applications. Their web environment is deployed by different purpose applications like we have different applications for social surfing, cloud storage, emails, online marketing, etc. With all these bundles of online applications, online frauds and online crimes are increasing swiftly [21]. Many online actions are punishable by a court of law. To justify any case or prove any complaint experts fetch data present in digital devices known as digital evidence. While the process of investigation, experts have to collect and analyze many devices and their data, which makes this process difficult. The process of investigation may vary from device to device and case to case. One process used for investigation in one case for one device may completely be different from other cases and another device. Hence it is really hard to find a compatible investigation process for digital devices. Cloud storage and online disk, where users can store their data are generation problems during the investigation. Forensic devices are getting advanced day by day, but anti-forensics development has obstructed the path of digital forensic investigation [22].

The proposed framework is about crime detection and control. The primary goal of crime detection and diminution in digital forensics (CD3F) is to help the investigator explain how specific digital evidence is discovered on a device. Despite the existence of numerous frameworks in the current literature, the CD3F methods and nomenclature have yet to be properly standardized. Attempts to standardize computer investigative process frameworks in the past appear to have not been fully successful for a variety of reasons. The authors' main reason for failing is that they used their own vocabulary instead of seeking to discover the most common language that can be accepted universally by digital forensic investigators. Fig. 1 is block representation of the proposed digital forensic guidance model (CD3F) and the detailed process flow of the same is explained further. The suggested model's first phase, "Forensic analysis," examines case reporting and determines whether a digital forensic investigation is required for the same. The "preparation" phase follows, which focuses on the initial preparation for the case by examining the required set of acts and getting a search warrant or other necessary authority. The "examination" phase follows, which focuses on searching for evidence at the crime scene or in a location where evidence might be found. The "identification" and "collection" phases were devoted to locating and gathering prospective evidence shards. The "analysis" phase focuses on reviewing and analyzing the acquired material in order to collect evidence that can aid in the "analysis" phase. The final phase is "forensic reporting," in which the investigator reports the evidence discovered throughout the investigation. The proposed digital forensic model has the following phases.

#### A. Forensic Request

This is the first stage of the digital forensic process when incidence is reported and higher authorities hand over the case to digital forensic experts. The forensic request phase of a

digital investigation begins when an incident is detected by either internal events such as an intrusion detection system or external events such as a crime reported to the police. The occurrence must be confirmed or denied after it has been discovered and reported. However, once the incident is confirmed, the investigators must be notified so that the first response can begin. The digital forensic investigation may be of these two kinds.

a) *Public Digital Forensic Investigation:* In the Public digital forensic investigation method, government body is responsible for the investigation of the registered case request. During this investigation of a criminal case, the investigator must have a sound knowledge of local city/town, state, country, and country laws related to the criminal case and all cyber-crime laws and all standard legal processes of investigation [23].

b) *Private Or Corporate Digital Forensic Investigation:* Whereas private investigation deals with private lawyers and companies who look after legal issues and that particular company policy violation. In such an investigation, the investigation must take care of business and should not let it suffer and the investigator should interrupt company employees to the minimal during the investigation. Investigators cannot seize the evidence, rather than seize it, they acquire memory images and allow the system to go back to work [23].

### B. Preparation

The next phase is to get prepared with the tools and methods which will be used further for the process and if required, train and build the forensic expert team for the investment. It is also suggested in this phase that if a search warrant or other documents are required then try to obtain in the initial level of this model. These documents will help in the further investigation process. Basically, in this phase investigator identifies the incidence and calculates possible risk assessment. Investigators determine which software and what kind of hardware will be required for investigation. He/ she will also try to define if any specific tool will be required for the investigation process to fetch information that will be further used as 'evidence'.

### C. Examination

In this phase of the investigation, the investigator will lock the crime scene's physical environment. The digital forensic expert will try to secure all correlated logs, data, and volatile evidence like laptops, mobile phones, and hardware and he/she will also ensure that condition of electronic devices won't get altered by any means. Investigators aim to examine and identify similar past investigations in this phase. If they find one, they study it and follow the footprint of that investigation, which can help them during the investigation from a secure physical location. Investigators also verify the extent of the damage/impact of the incident and ensure that non-digital evidence such as fingerprints are protected. Observe and document the physical scene, device positions, device locations relative to one another, and device conditions, including power status.

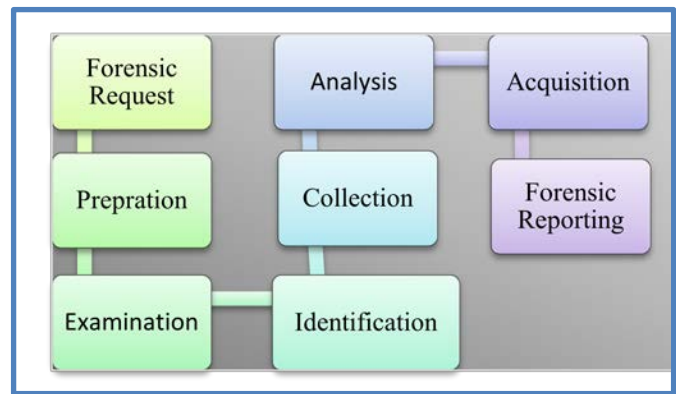


Fig. 1. The Proposed Digital Forensic Guidance Model.

### D. Identification & Collection

The next phase in this model explains about identification and collection of digital evidence after performing examination of evidence around the crime scene. After the collection of evidence from the crime scene, an important and necessary step is to preserve all evidence so that they can utilize it further for examination (if required) and their integrity will remain constant. Detailed information regarding the evidence will be included into the evidence gathering form during this step. If electronic evidence is being used, a deeper inquiry into the device's volatile data will be required. If volatile data is needed, execute a live acquisition of volatile data first, then verify if non-volatile data is needed, and then perform a live acquisition of non-volatile data. Make a duplicate copy of the data you just got and double-check it. This phase will involve labelling the evidence discovered, as well as packing and transporting the evidence. The evidence is then placed in a legal custody room with labels and security measurements. Maintaining and preserving the chain of custody is essential.

### E. Analysis

This phase focuses on the process of examining digital evidence that has been discovered. During the analysis, the examiner assembled the evidence to gather information, and after reviewing the information, the examiner may develop certain conclusions for final reporting. During this phase, the investigator gathers unprocessed data and devices from the linked investigation and compares them to the condition of the device obtained through the related investigation. Investigator Extracts data from gathered devices using physical, logical, and dead acquisition methods.

### F. Acquisition

This is a critical stage in the digital forensic investigation process. In this case, digital forensic professionals will attempt to gather all unprocessed data and devices from the investigation. After successfully collecting physical and logical data, he will make duplicate copies of all collected data so that the original data is not affected, and only the investigator will execute all operations and analyses on the copied data. During this phase, the investigator compares duplicate data to the original in terms of timestamp and reconstructs the data taken from devices. Only the investigator selects an analysis approach from a list of options, such as data hiding analysis, log analysis, timeframe analysis, application and file analysis,



and so on. Finally, reconstructs the chronology of crimes in order to provide a clear picture and discover missing links in order to locate relevant evidence.

#### G. Forensic Reporting

This is the final phase of the digital forensic investigation framework, and it entails reporting and drawing conclusions from all previous phases, as well as all essential material. The conclusion will be drawn and reported to the court of law. This is the final phase of the CD3F architecture; investigators prepare a detailed report that can be understood by laypeople, choose the target audience, gather evidence, and maintain the chain of custody throughout this phase. Closer documents will be provided by the investigator, along with the time and date of release, as well as to whom and by whom they were released. This evidence will be presented in a court of law to assist in the resolution of the case.

There are primarily eight steps in the proposed digital forensic guiding model that describe the investigative process. This digital forensic investigation technique entails obtaining digital data for inspection in order to use the information discovered as evidence in reopened cases. The type and format of this digital record can vary. Smartphone data, a list of all phone calls made, desktop files, recorded video and audio files, a bit of signal strength from a mobile SIM station's base station, all electronic mail chats, installed and attacked virus, and so on [24]. Once investigators have obtained these records, the most important next step is to create copies of all evidence, which will then be examined and analyzed so that the integrity of the original evidence is not compromised and no issues are raised about its integrity.

#### IV. THE PROPOSED DIGITAL FORENSIC GUIDANCE FRAMEWORK DETAILS

This model is influenced by previous existing models as well as some physical forensic models so that model can encounter the challenges from the electronic evidence to make them admissible in a court of law. This framework has two major contributions. First, it describes a framework that is trying to cover previous events and the state of electronic devices at the primitive and abstract level of investigation. Second, it provides detailed steps for each phase, so that it can use as a reference investigation framework. The following goals were used to define the model:

- The framework is designed based on the theoretical foundations of digital forensic investigation so that current and future work in digital forensics science can be used.
- The framework must be general with high opinion based on the technology being investigated so that this theory could be applied to upcoming as well as existing technologies.
- The framework must be adept at events, supporting systems, and storage locations at arbitrary levels of abstraction so that complex systems can be represented.

- The framework designed as per it is capable of describing past events and states so that all electronic evidence can be represented.

After obtaining a forensic request for a registered or reported case digital forensic investigation progress gets triggered. Here is the description of the phases of the proposed investigation model in brief.

#### A. Preparation Phase

The initial understanding of the problem, as well as the appropriate tools, are all part of the preparation phase. This step is used to get authorization and approval, as well as a search warrant and legal notification to people who have expressed concern, before developing a suitable plan. Detailed explanation is given in Fig. 2. Here's a rundown of the steps involved in the planning phase.

- 1) Identify or detect incidence and possible risk assessment of the reported case.
- 2) Actuate Computer Emergency Response Team (CERT) divide preliminary assignment and maintain legal activity coordinated plan before arriving at the crime scene.
- 3) If required obtain a search warrant and permission from concerned authorities.
- 4) Formulate paperwork according to the requirement of the case and gather all needed requirements and identify requirements.
- 5) Develop an onsite plan which includes policies and individual responsibilities.
- 6) Select approach and strategy for collection, preservation, examination, and analysis of evidence.
- 7) Have any information about the suspected operation system.
- 8) Determine the kind of software and hardware for investigation. Specific tool, accumulate evidence collection, and packaging equipment and materials.
- 9) If more information does not require further processing. Then move to the next step of digital forensic investigation.

#### B. Examination Phase

The second phase focuses on protecting the crime scene from illegal entry and preventing contamination of the evidence. An early investigation by the investigators to assess the crime scene, identify potential sources of evidence, and devise a search strategy. This phase entails photography, sketching, and crime-scene mapping, as well as the adequate recording of both physical and digital crime scenes. Detailed explanation is given in Fig. 3. The brief of the examination phase is as follows -

- 1) The most first step is to secure the crime scene's physical environment & secure all correlated logs, data, and volatile evidence. Laptops, hardware, and secure narrative description. Don't alter the condition of electronic devices.

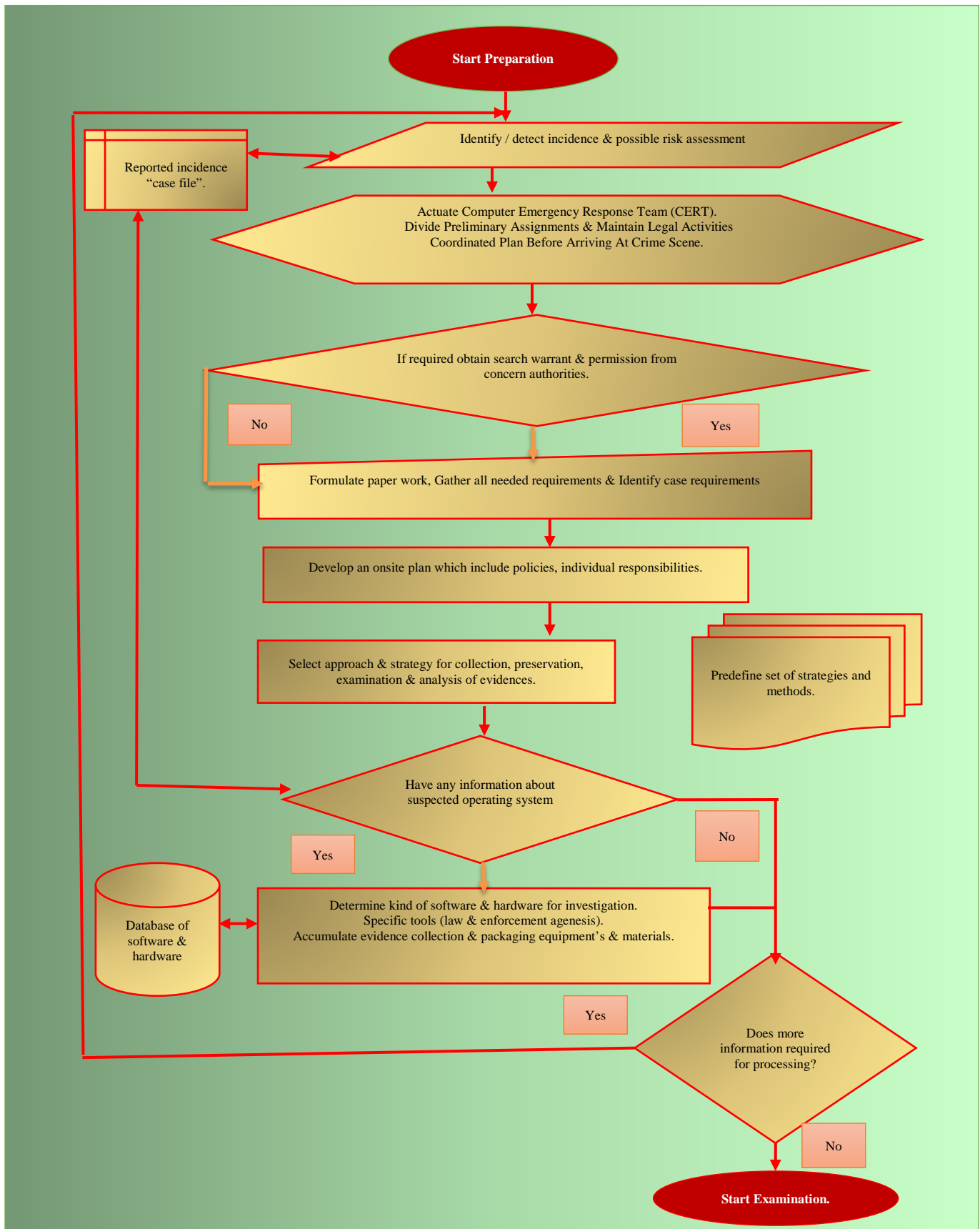


Fig. 2. Dataflow Diagram for Preparation Phase Digital Forensic Investigation Model.

2) Try to analyze and find similar previous investigations; if find one then study similar investigation & follow the footprint of that investigation that can help during the investigation from the secure physical environment in step 1.

3) Place labels over all the drive slots and power connections and take preliminary photographs of the crime scene.

4) Select narration technique (written, audio or video) to delineate the search area and detect unauthorized activity and report it.

5) Validate the damage/ impact of incidence and ensure the protection of non-digital evidence like fingerprints.

6) Evaluate whether any movement appears in evidence, determine devices on the network and make a complete evolution sheet.

7) Observe & document the physical scene, the position of devices, the location of devices relative to each other, and the condition of devices including power status.

8) Take written notes on what appears on the screen, take snapshots of the screen, and the active program should be videotaped.

9) Take photographs before and after examination of evidence. Label properly each evidence.

10) Maintain and seize evidence log that includes a brief description and photographic log. Prepare a chain of evidence.

11) Start Identification & collection.

### C. Start Identification & Collection Phase

In the identification phase investigator needs to disable all other possible communication methods for the devices. Some communication technologies, such as WiFi or Bluetooth, may be enabled even if the device appears to be turned off. This may result in the overwriting of existing data, so such scenarios should be avoided. In evidence, both volatile and non-volatile evidence could be present. To preserve its integrity, the required precautions must be performed. Detailed explanation is given in Fig. 4. A brief of the identification and collection is given below-

1) Firstly, check whether evidence to be collected are physical or electronic?

2) If evidence is physical then apply the tag on an identified object as evidence like removable media, cables, publications, and all computers. Or if the evidence is electronic then check whether the device is running or not.

3) Fill evidence collection form with detailed information about the evidence.

4) If electronic evidence is running, then checking for volatile data of the device will require further investigation.

5) If volatile data is required, then perform live acquisition of volatile data then check non-volatile data is required then perform live acquisition of non-volatile data.

6) Check whether found device data is stable. If yes, then remove the power source whether battery or main switch. If no, then perform a normal system shutdown.

7) Decide the most appropriate way to acquire data and then acquire data from the device.

8) Make a duplicate copy of the acquired data and verify.

9) Check whether all required data has been acquired. If yes, then seize found device.

10) Record and return the connection of the device. Label the evidence found then pack and transport the evidence.

11) Store the evidence in a legal custody room with labels and security measurements. Maintain and preserve the chain of custody.

### D. Analysis and Acquisition

Examining the content of the acquired evidence and extracting information for presentation in court is what a forensics specialist does. This consists of both volatile and non-volatile data. The acquisition is more of a technical evaluation undertaken by the investigation team based on the findings of the digital evidence inspection and the reconstruction of event data. Detailed explanation is given in Fig. 5. The brief of this phase is given below-

1) Collect unprocessed data and devices from the related investigation.

2) Identify operating systems used in incidence & choose data extraction techniques for examination & analysis of evidence.

3) Check documents obtained by related investigation of the condition of the device.

4) Perform physical, logical extraction, and dead acquisition on data from collected devices.

5) Make duplicate copies of all acquired data from electronic devices.

6) Authenticate duplicate data with the original one in their timestamp.

7) Reconstruct the extracted data from devices.

8) Choose the analysis technique - Data hiding analysis, log analysis, Timeframe analysis, Application & file analysis, etc.

9) Reconstruct the sequence of crimes to produce a clear picture & try identifying missing links.

10) Compare acquired evidence with proven facts and with physical forensic results.

11) Documentation & Preserve chain of custody in storage.

12) Store evidence in a secure custody room.

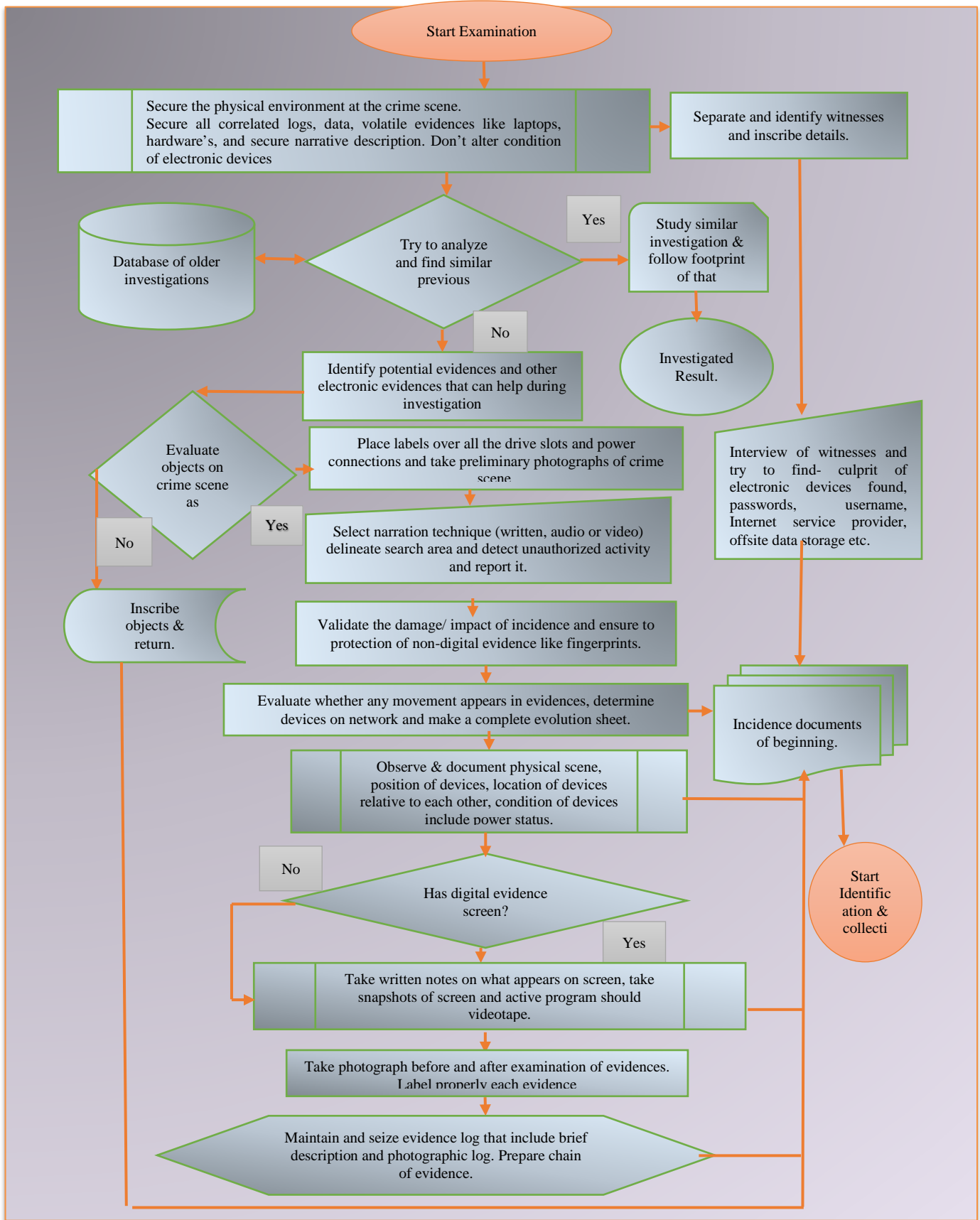


Fig. 3. Dataflow Diagram for Examination Phase of Digital Forensic Investigation Model.

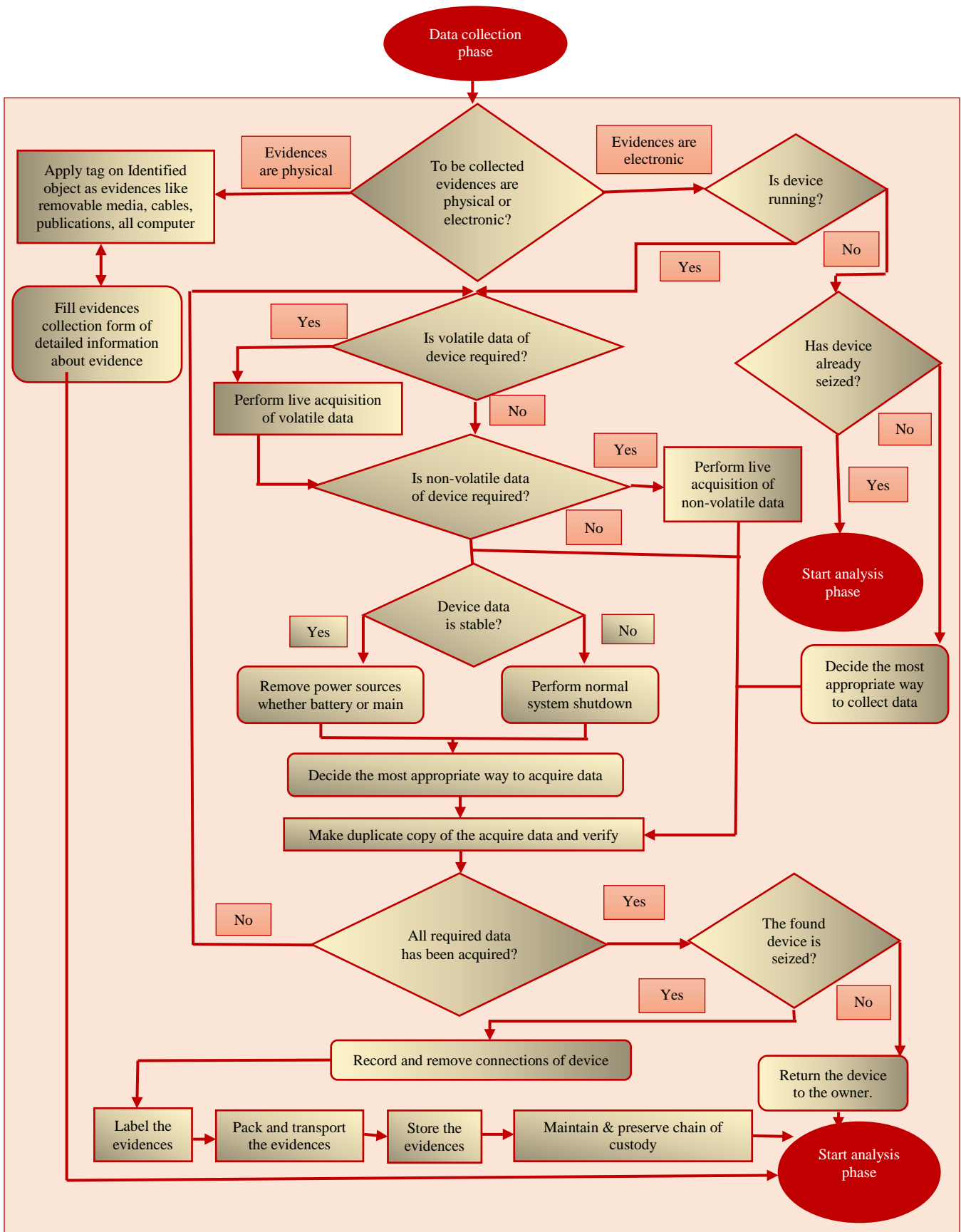


Fig. 4. Dataflow Diagram for Collection Phase of Digital Forensic Investigation Model.

### *E. Reporting & Presentation*

A report containing a full overview of the various procedures done during the investigation of evidence and System (Organization of potential evidence) and the conclusion reached is presented to the proper authorities [48]. This model was a basic workflow without any validation and verification. The model was designed to help individual investigators in organizing, but this was a very complex puzzled workflow. Presentation phase: When a crime is committed, it is presented to a court of law, and when an event occurs, it is presented in court. At the conclusion of the investigation, an evaluation is conducted, and the results are utilized to update or repair any shortcomings discovered during the inquiry Detailed

explanation is given in Fig. 5. A brief of this phase is given below-

- 1) Write a comprehensive report which can be understood by the layman as well.
- 2) Determine the target audience and put together evidence and preserve the chain of custody.
- 3) Present evidence according to rules of the law enforcement.
- 4) Preserve evidence for further requirements.
- 5) Handover closer Documents, with time & date of release, to whom & by whom released.



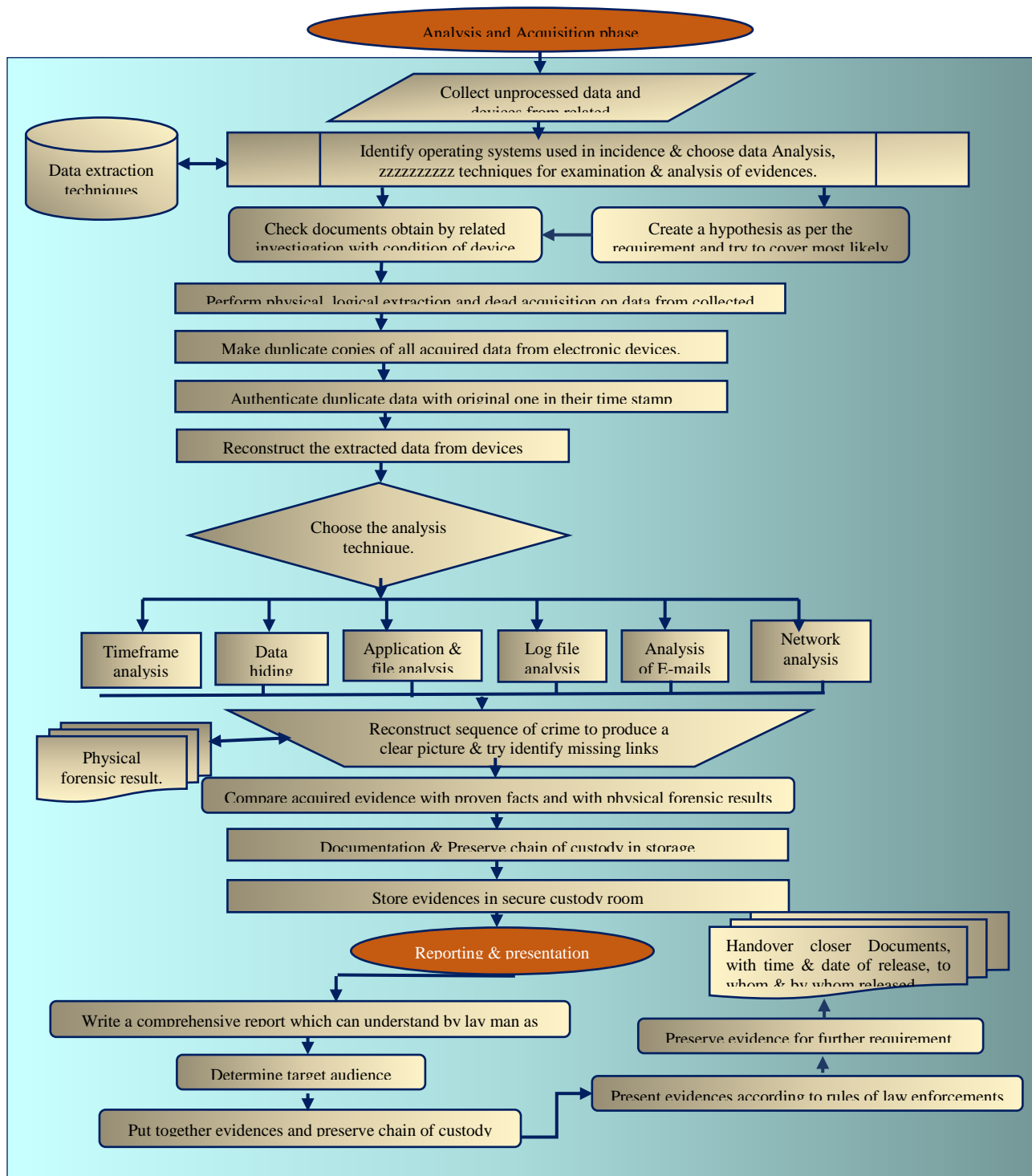


Fig. 5. Dataflow Diagram for Analysis, Acquisition, Reporting, and Presentation Phase of Digital Forensic Investigation Model.

## V. DISCUSSION

Dimpe and Kogeda [47] examined the earlier proposed models and found the integrity of evidence can be preserved by documentation of each action performed during the investigation. To serve this purpose they proposed a generic investigation framework. This framework is focused on the integrity of evidence, while collecting evidence in the collection phase this phase will be divided into sub-phases –

acquisition, transport, and storage. The author claimed that they explained requirements during the digital forensic investigation to make investigation work easy and explained standers. But setting investigation standards cannot solve all problems and challenges faced by investigation officers. Dimpe and Kogeda analyze the need for documentation for the integrity of evidence, but their main focus was on standardizing phases of investigation and developing skills for

the investigator. But even we have national and international courses and training systems for professionals. If through documentation only the investigator can save the integrity of evidence, then we will not be facing any of such problems yet. Hence, we proposed standards and an easy framework by analyzing the earlier proposed one that accommodates everything which could be considered and followed in the investigation process.

Bulbul, Yavuzcan, and Ozel suggested a model named “Digital forensics: An Analytical Crime Scene Procedure Model (ACSPM)” which focused on procedures of crime scenes [48]. This multi-stage model tries to offer a series of evidence collection procedures and multiple tasks for the crime scene to support the investigation process. This proposed model is having some new suggested tasks like Crime scene examination, Evidence search, Potential evidence acquisition, System assurance and management. This model was incapable to focus on all electronic devices and all electronic pieces of evidence which needed to be collected during the investigation.

Ohaeri and Esiefarienhe proposed a digital forensic model for network security management and information system. In this model digital forensic investigation stages were implemented as a security mechanism. The basic thought behind this model was to provide a detailed knowledge of the digital forensic technology practices in institutions, organizations, or companies [49]. The model is claimed to ensure uniqueness and effectiveness while succeeding to provide adequate, reliable, and effective security. But this model is not successfully achieving its goal of real-time dealing with the investigation. It is just like a traditional search method for investigation. According to the author, data integrity is justified while keeping the laws and rules of digital evidence but the model seems to fail in proper documentation of evidence. This can be raised finger on the integrity of evidence.

Graeme Horsman proposed DERDS framework to support the digital forensic investigation. This model serves with

logical decisions and the investigator those are experienced but lack confidence. This guidance model is a process flow for making the right judgment with the help of found digital evidence. The DERDS framework delivers three corridors-inferences, assumptions, or conclusions, for an investigator to test and search for the consistency of digital investigative [50]. But this model capacity is depending on the ability of the investigator or researcher. DERDS framework always needs a doorkeeper for proceeding further in the investigation and for key decisions during the investigation process.

Author suggested a digital forensic investigation framework in our proposal, hoping to provide an optimistic method to researching cyber-attacks. This structure is mostly made up of four-fold. First, it aids in the digital forensic investigation model's preparation phase. Second, is the digital forensic inquiry model's examination step. The digital forensic investigation model's third phase, is collection. The fourth phase of the digital forensic investigation process is the examination, acquisition, reporting, and presentation. We proposed a simple and standard methodology that includes accurate documentation at each stage and attempts to cover all parts of the investigation.

## VI. CRITICAL EVALUATION OF THE PROPOSED FRAMEWORK

The proposed framework is based on a holistic approach which is able to focus and combine all aspects of digital forensic investigation. The processes provided in the proposed framework is vital in investigation and provide more advantages. The proposed framework tries to cover all aspects of the investigation process and all predefined framework processes which show that this framework is enough comprehensive to cover whole aspects of the investigation. One of the important benefits of the proposed framework is fetching out potential evidence forensically to improve admissibility in a court of law. Table I is showing a comparison of the CD3F framework with some pre-existing frameworks.

TABLE I. COMPARISON OF THE PROPOSED FRAMEWORK FROM PRE-EXISTING FRAMEWORKS

	Framework	Year	Contribution	Loophole	Comparison from the proposed framework
[1]	Systematic Digital forensic Investigation Model [29]	2011	Model work for dynamic evidence & reconstruct events.	The process is similar to old process like only the terms used are different.	Proposed framework provide a less complex investigation path way to the investigator as well as compatible with the advance technology.
[2]	Integrated Digital Investigation Process Model[30]	2011	Identifies the need for interaction with resources in right way.	Proposed Interaction tool needs proper training & patience.	Proposed framework can be used in any digital forensic investigation with only basic training of investigator.
[3]	Generic Digital forensic Framework[25]	2013	Set standard requirement for digital forensic investigation.	Explained standard does not satisfy promise.	Framework explains set of slandered required in each phase.
[4]	An analytical crime scene Proceeding Model (ACSPM)[26]	2018	Talk about management of digital evidence & crime scene investigation	Only focused on crime scene procedure.	Proposed framework is focused on every aspect of investigation like crime scene investigation, evidence collection, analysis, lab examination etc.
[5]	Digital Evidence Reporting and Decision Support (DERDS) framework[27]	2019	Guidance model for the investigator, when to report findings to minimize unsafe disclosure of evidence	Not 100% error-free & may not agree to report all evidence so that evidence may get lost.	Clearly report all evidence so that it requires reinvestigation, evidence recall is unbiased.

## VII. SIGNIFICANCE OF STUDY

The framework highlighted certain phase commonalities that may be regrouped to make the framework more logical. For example, Survey and Recognition could be part of Preparation, Documenting the Crime Scene could be part of Securing the Crime Scene, and Communication Shielding could be part of Securing the Crime Scene because these two independent phases in this model are actually part of Securing the Crime Scene. It's also possible to mix examination with analysis. These phrases were employed as different activities in the model, although their definitions are not just comparable, they also complement one other, which can lead to confusion if they are separated. The following are some advantages of the proposed framework.

- In the proposed framework, a standardized process is used. This makes higher chances of extracting the potential evidence during the investigation.
- The proposed framework is based on a holistic approach and the framework is also able to incorporate the existing frameworks, and thereby this framework could be used as a harmonized model during IoT environment investigation.
- Throughout the whole process of the proposed framework integrity of collected potential evidence is preserved.
- The proposed framework provides a less complex investigation pathway to the investigator as well as is compatible with advanced technology.
- From a reinvestigation point of view, all useful information and all extracted evidence are preserved digitally.

In the proposed framework, all processes can be executed continuously and also insuring the evidence admissibility in a court of law. The authors have involved the concurrent processes in the proposed framework as per guided in ISO/IEC 27043: 2015 standards. A comparison with existing models by the table is also been done in this paper which will further bring out the efficiency of the proposed framework.

## VIII. CONCLUSION AND FUTURE WORK

The author provided a framework for digital forensics in this study. As technology advances, so does the number of incidents of cybercrime. Over previous models, the suggested framework's level of comprehensive processes for each phase independently offers unique benefits. This is a framework that consists of steps that the investigator must follow during the investigation. It is a generic/universal framework that is not dependent on technology or limited to a set of tools. As a result, it will not be constrained by current technologies. The suggested framework is technology-neutral, it may be used in a variety of research platforms and scenarios. This framework could also be utilized in a variety of digital forensics cases. This study will help the different stake holders to detect the crime at a very early stage (as explained in the examination phase, step 2) by following the old recorder investigation footprint. This detection at an early stage can reduce the

detection time of the crime and hence can reduce the further process time in the digital forensics investigation.

Since CD3F only includes activities for on-site plane and lab work, the study described in this paper is clearly lacking. In order to maximize the process of its continued development, the proposed CD3F would need to be applied to a number of case studies with a methodical approach. It is acknowledged that the work still has some limitations in terms of future work. Despite the fact that the CD3F has previously been reviewed by a few knowledgeable experts in the area. As a result, as part of a bigger study, future work should involve a more thorough investigation by digital forensic investigators. The forensic laboratory's typical examination and analysis might also be covered by the CD3F in the upcoming work, which would help to ensure the work in more controlled environment.

## REFERENCES

- [1] Mousa, M. Karabatak and T. Mustafa, "Database Security Threats and Challenges," 2020 8th International Symposium on Digital Forensics and Security (ISDFS), 2020, pp. 1-5, doi: 10.1109/ISDFS49300.2020.9116436.
- [2] The Hindu business (17 January 2018), India lagging in cyber security awareness. Available at : [ <https://www.thehindubusinessline.com/info-tech/india-lagging-in-cyber-security-awareness/article9046626.ece> ] access on – 2 June 2021
- [3] P. Čisar and Sanja Maravić Čisar, "Methodological frameworks of digital forensics," 2011 IEEE 9th International Symposium on Intelligent Systems and Informatics, 2011, pp. 343-347, doi: 10.1109/SISY.2011.6034350.
- [4] Goel S. 2020. National cyber security strategy and the emergence of strong digital borders. *Connections: The Quarterly Journal* 19(1):73–86 DOI 10.11610/Connections.19.1.07.
- [5] Reid R, Van Niekerk J. 2014. From information security to cyber security cultures—information security for South Africa. *Piscataway: IEEE*, 1–7.
- [6] CISA. 2020. Critical infrastructure sectors. Available at <https://www.cisa.gov/critical-infrastructuresectors> (access on April 11, 2022)
- [7] Check Point Security Report. 2020. Check point research. Available at <https://research.checkpoint.com/>
- [8] Wang D, Wang X, Zhang Y, Jin L. 2019. Detection of power grid disturbances and cyber-attacks based on machine learning. *Journal of Information Security and Applications* 46(27):42–52 DOI 10.1016/j.jisaa.2019.02.008.
- [9] Zolanvari M, Teixeira MA, Gupta L, Khan KM, Jain R. 2019. Machine learning-based network vulnerability analysis of industrial Internet of Things. *IEEE Internet of Things Journal* 6(4):6822–6834 DOI 10.1109/JIOT.2019.2912022.
- [10] Canbek G, Sagiroglu Ş, Temizel TT. 2018. New techniques in profiling big datasets for machine learning with a concise review of android mobile malware datasets. In: 2018 International Congress on Big Data, Deep Learning and Fighting Cyber Terrorism (IBIGDELFT). 117–121.
- [11] Moorthy RS, Pabitha P. 2020. Optimal detection of phishing attack using SCA based K-NN. *Procedia Computer Science* 171(5):1716–1725 DOI 10.1016/j.procs.2020.04.184.
- [12] Ngejane CH, Mabuza-Hocquet G, Eloff JH, Lefophane S. 2018. Mitigating online sexual grooming cybercrime on social media using machine learning: a desktop survey. In: 2018 International Conference on Advances in Big Data, Computing and Data Communication Systems (icABCD). Piscataway: IEEE, 1–6.
- [13] Gurjar M, Naik P, Mujumdar G, Vaidya T. 2018. Stock market prediction using ANN. *International Research Journal of Engineering and Technology* 5:2758–2761.
- [14] Wheeler AP, Steenbeek W. 2020. Mapping the risk terrain for crime using machine learning. Epub ahead of print 24 April 2020. *Journal of Quantitative Criminology* DOI 10.1007/s10940-020-09457-7.

- [15] Zufadhilah M, Prayudi Y, Riadi I. 2016. Cyber profiling using log analysis and k-means clustering. *International Journal of Advanced Computer Science and Applications* 7(7):430–435 DOI 10.14569/IJACSA.2016.070759.
- [16] Biswas AA, Basak S. 2019. Forecasting the trends and patterns of crime in Bangladesh using machine learning model. In: 2nd International Conference on Intelligent Communication and Computational Techniques (ICCT). Piscataway: IEEE, 114–118.
- [17] Bharathi ST, Indrani B, Prabakar MA. 2017. A supervised learning approach for criminal identification using similarity measures and K-Medoids clustering. In: ICICICT. Piscataway: IEEE, 646–653.
- [18] Lin YL, Chen TY, Yu LC. 2017. Using machine learning to assist crime prevention. In: 6th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI). Piscataway: IEEE, 1029–1030.
- [19] M. Reith, C. Carr, and G. Gunsch, "An examination of digital forensic models" *international journal of digital evidence*, 2002.
- [20] S. A. Ali, S. Memon, and F. Sahito, "Challenges and solutions in cloud forensics" in *Proceedings of the 2018 2nd International Conference on Cloud and Big Data Computing*. ACM, 2018, pp. 6–10
- [21] S. Raghavan, "Digital forensic research: current state of the art", *ICT Transactions on ICT*, vol. 1, no. 1, pp. 91–114, 2013.
- [22] Al-Masri, E., Bai, Y., & Li, J. (2018). "A Fog-Based Digital Forensics Investigation Framework for IoT Systems". 2018 IEEE International Conference on Smart Cloud (SmartCloud). doi:10.1109/smartcloud.2018.00040
- [23] Bonomi, F., Milito, R., Zhu, J. and Addepalli, S., 2012, August. Fog computing and its role in the internet of things. In *Proceedings of the first edition of the MCC workshop on Mobile cloud computing* (pp.13-16). ACM.
- [24] Islam, M. J., Mahin, M., Khatun, A., Debnath, B. C., & Kabir, S. (2019). Digital Forensic Investigation Framework for Internet of Things (IoT): A Comprehensive Approach. 2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT). doi:10.1109/icasert.2019.8934707
- [25] Luan, T.H., Gao, L., Li, Z., Xiang, Y., Wei, G. and Sun, L., 2015. Fog computing: Focusing on mobile users at the edge. *arXiv preprint arXiv:1502.01815*.
- [26] Kaur Chahal J, Bhandari A, Behal S. 2019. Distributed Denial of service attacks: a threat or challenge. *New Review of Information Networking* 24(1):31–103 DOI 10.1080/13614576.2019.1611468.
- [27] Sahingoz OK, Buber E, Demir O, Diri B. 2019. Machine learning based phishing detection from URLs. *Expert Systems with Applications* 117(4):345–357 DOI 10.1016/j.eswa.2018.09.029.
- [28] Biju JM, Gopal N, Prakash AJ. 2019. Cyber attacks and its different types. *International Research Journal of Engineering and Technology* 6(3):4849–4852
- [29] Mitnick KD, Simon WL. 2009. *The art of intrusion: the real stories behind the exploits of hackers, intruders and deceivers*. Hoboken: John Wiley & Sons
- [30] Breda F, Barbosa H, Morais T. 2017. Social engineering and cyber security. *International Technology, Education and Development Conference* 3(3):106–108
- [31] Kagita MK, Thilakarathne N, Gadekallu TR, Maddikunta PKR, Singh S. 2020. A review on cyber crimes on the Internet of Things. *arXiv*. Available at <http://arxiv.org/abs/2009.05708>
- [32] Rewari S, Singh W. 2017. Systematic review of crime data analytics. In: *International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI)*. Piscataway: IEEE, 3042–3045
- [33] Hassan M, Rahman MZ. 2017. Crime news analysis: location and story detection. In: 20th International Conference of Computer and Information Technology (ICCIT). Piscataway: IEEE, 1–6
- [34] Zhao X, Tang J. 2017. Exploring transfer learning for crime prediction. In: *IEEE International Conference on Data Mining Workshops (ICDMW)*. Piscataway: IEEE, 1158–1159
- [35] Vineeth KRS, Pandey A, Pradhan T. 2016. A novel approach for intelligent crime pattern discovery and prediction. In: *International Conference on Advanced Communication Control and Computing Technologies (ICACCCT)*. Piscataway: IEEE, 531–538.
- [36] Feng M, Zheng J, Han Y, Ren J, Liu Q. 2018. Big data analytics and mining for crime data analysis, visualization and prediction. In: *International Conference on Brain Inspired Cognitive Systems*. Cham: Springer, 605–614.
- [37] Bharati A, Sarvanaguru RAK. 2018. Crime prediction and analysis using machine learning. *International Research Journal of Engineering and Technology* 5(9):1037–1042
- [38] Kim S, Joshi P, Kalsi PS, Taheri P. 2018. Crime analysis through machine learning. In: 2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON). Piscataway: IEEE, 415–420.
- [39] Chandrasekar A, Raj AS, Kumar P. 2015. Crime prediction and classification in San Francisco City. Available at [http://cs229.stanford.edu/proj2015/228\\_report.pdf](http://cs229.stanford.edu/proj2015/228_report.pdf).
- [40] Alves LGA, Ribeiro HV, Rodrigues FA. 2018. Crime prediction through urban metrics and statistical learning. *Physica A: Statistical Mechanics and its Applications* 505:435–443 DOI 10.1016/j.physa.2018.03.084.
- [41] Kumar A, Verma A, Shinde G, Sukhdeve Y, Lal N. 2020. Crime prediction using K-nearest neighboring algorithm. In: 2020 International Conference on Emerging Trends in Information Technology and Engineering. Piscataway: IEEE, 1–4.
- [42] Jang-Jaccard J, Nepal S. 2014. A survey of emerging threats in cybersecurity. *Journal of Computer and System Sciences* 80(5):973–993 DOI 10.1016/j.jcss.2014.02.005.
- [43] Verma D, Yarlagadda R, Gartner SS, Felmler D. 2019. Understanding patterns of terrorism in india (2007–2017) using artificial intelligence machine learning. *International Journal of Technology, Knowledge, and Society* 15(4):23–39 DOI 10.18848/1832-3669/CGP/v15i04/23-39
- [44] Arora T, Sharma M, Khatri SK. 2019. Detection of cyber crime on social media using random forest algorithm. In: 2nd International Conference on Power Energy, Environment and Intelligent Control (PEEIC). Piscataway: IEEE, 47–51
- [45] Ghankutkar S, Sarkar N, Gajbhiye P, Yadav S, Kalbande D, Bakereywal N. 2019. Modelling machine learning for analysing crime news. In: 2019 International Conference on Advances in Computing, Communication and Control (ICAC3). 1–5
- [46] Ch R, Gadekallu TR, Abidi MH, Al-Ahmari A. 2020. Computational system to classify cyber crime offenses using machine learning. *Sustainability* 12(10):4087 DOI 10.3390/su12104087.
- [47] P. M. Dimpe and O. P. Kogeda, "Generic Digital Forensic Requirements," 2018 Open Innovations Conference (OI), 2018, pp. 240-245, doi: 10.1109/OI.2018.8535924.
- [48] Bulbul, H.I., Yavuzcan, H.G., and Ozel, M., Digital forensics: An Analytical Crime Scene Procedure Model (ACSPM), *Forensic Science International*, Volume 233, Issues 1–3, (2013) Pages 244-256, ISSN 0379-0738, doi:10.1016/j.forsciint.2013.09.007.
- [49] I. U. Ohaeri and B. M. Esiefarienhe, "Digital Forensic Process Model for Information System and Network Security Management," 2018 International Conference on Computational Science and Computational Intelligence (CSCI), 2018, pp. 65-70, doi: 10.1109/CSCI46756.2018.00020.
- [50] Horsman G. ,Formalising investigative decision making in digital forensics: Proposing the Digital Evidence Reporting and Decision Support (DERDS) framework, *Digital Investigation*, Volume 28, (2019), Pages 146-151, ISSN 1742-2876, doi. :10.1016/j.diin.2019.01.007.

# Deep Learning and Classification Algorithms for COVID-19 Detection

Mr. Mohammed Sidheeque<sup>1</sup>  
Research scholar, School of  
Computer Science  
Engineering & Applications  
Bharathidasan University  
Tamil Nādu, India

Dr. P. Sumathy<sup>2</sup>  
Assistant Professor, School of  
Computer Science  
Engineering & Applications  
Bharathidasan University  
Tamil Nādu, India

Dr. Abdul Gafur. M<sup>3</sup>  
Principal  
Ilahia College of Engineering and  
Technology  
Ernakulam, Kerala  
India

**Abstract**—The imaging modalities of chest X-rays and computed tomography (CT) are commonly utilized to quickly and accurately diagnose COVID-19. Due to time and human error, it is exceedingly difficult to manually identify the infection using radio imaging. COVID-19 identification is being mechanized and improved with the use of artificial intelligence (AI) tools that have already showed promise. This study employs the following methodology: The chest footage was pre-processed by setting equalizing the histogram, sharpening it, and so on. The transformed chest images are then retrieved through shallow and high-level feature mapping over the backbone network. To further improve the classification performance of the convolutional neural network, the model uses self-attained mechanism through feature maps. Numerous simulations show that CT image classification and augmentation may be accomplished with higher efficiency and flexibility using the Inception-Resnet convolutional neural network than with traditional segmentation methods. The experiment illustrates the association between model accuracy, model loss, and epoch. Inception-statistical Resnet's measurement results are 98%, 91%, 91%.

**Keywords**—Deep Learning; COVID-19; classification; artificial intelligence

## I. INTRODUCTION

COVID-19, a new strain of the Coronavirus, was initially discovered in Wuhan, China in December of this year and has since spread rapidly over the globe [1, 2]. SARS-CoV-2, the virus that causes the disease, has infected millions of individuals throughout the globe. After infecting throat mucosa, COVID-19 may move to lungs through respiratory tract. In order to stop the spread of COVID-19 and expedite treatment, it is critical to quickly screen, identify, and isolate people who have the illness. Medical imaging, such as CXR and CT scans, have been shown to accurately diagnose COVID-19 infection and are now frequently utilized in disease screening [3–5]. However, owing to the disease's recent origins and resemblances to other respiratory conditions such as pneumonia, proper interpretation of findings via pictures presents various difficulties. Achieving a reliable diagnosis of COVID-19 is difficult and time consuming because of its complexity [6-10]. Only radiologists are qualified to conduct this work. Healthcare personnel all throughout the globe have a

difficult task as a result of this epidemic. Many patients' test findings must be analyzed over a period of time. In recent years, there has been a growth in the need for clinical assistance for COVID-19 patients' treatment [11,12]. In order to meet the needs of a large number of patients, image analysis on medical pictures combined with decision support systems may give an accurate and speedy diagnosis of illness. While radiologists can spend up to 10 minutes reviewing CT scans manually, decision support systems can do the same in less than a minute. Although COVID-19 may induce major organ malfunction, such as pneumonia and renal failure, it can also lead to mortality (Fleuren, Tonutti et al., 2021). As a result, early detection of COVID-19 patients is critical. The illness continues to spread since the patient is not isolated after the PCR test, which was used to diagnose COVID-19.

The rest of the paper is organized as follows; Section II describes about the related work; Section III describes proposed model; Section IV describes about the deep residual networks; Section V describes about results and discussions; Section VI describes about conclusions.

## II. RELATED WORK

Deep learning may be classified into unsupervised, supervised, and semi-supervised learning depending on the training dataset's labels. Images are all tagged via supervised learning, and the model is tuned using these image-label pairings. A probability score will be generated for each testing picture based on the model's optimum parameters [15]. It's possible to use unsupervised learning to discover patterns or hidden data structures without providing any labels to the model beforehand. In this case, the model learns input-output relationships from the labelled data and is enhanced by the unlabeled data, which contains more semantic and fine-grained information. Semi-supervised learning is a term for this sort of learning method [13]. We'll cover the most popular frameworks for each of these learning paradigms. The three broad methodologies mentioned here may be integrated with other learning standards for enhanced medical image processing performance. The general classification of the radiography image is represented in Fig. 1. The chest X-Ray images of various formats are represented in Fig. 2.

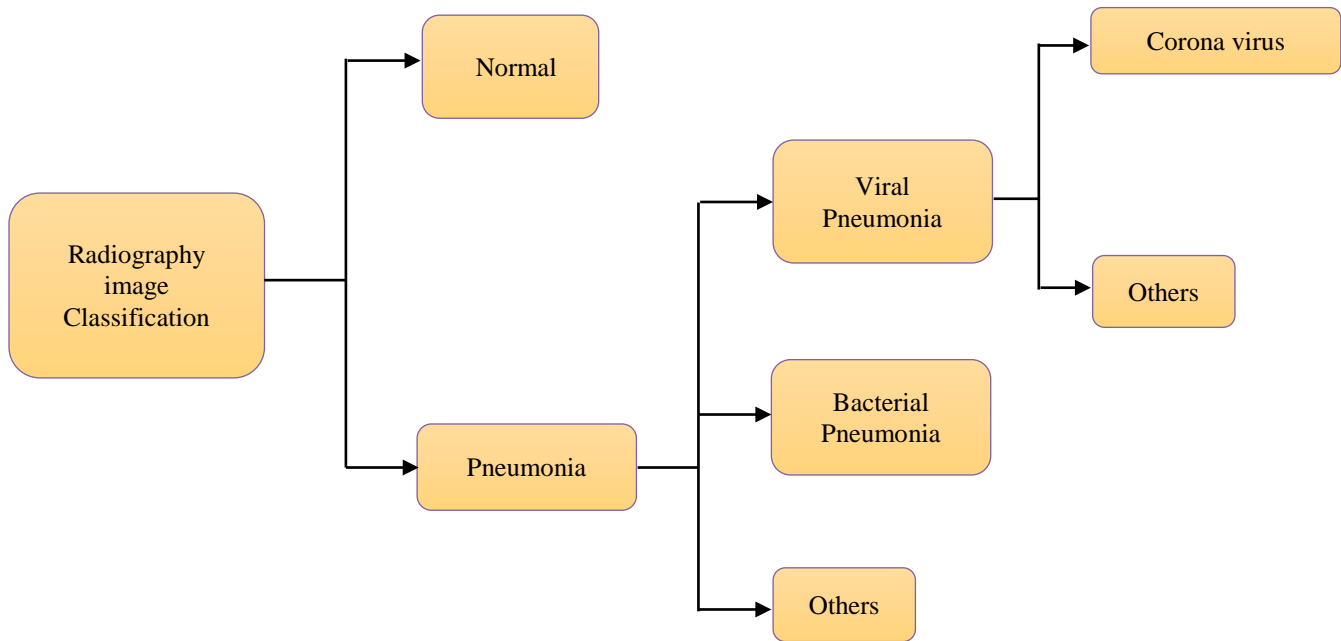


Fig. 1. General Classification of Radiography Images.

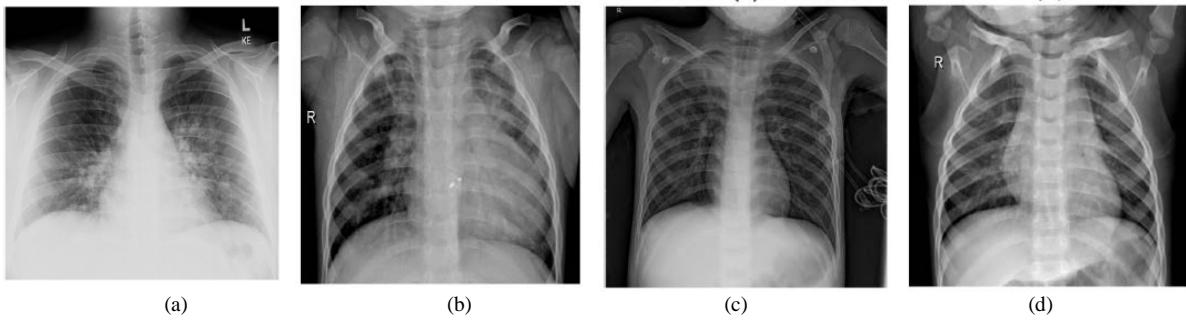


Fig. 2. Chest X ray Images (a) Covid 19, (b) Bacterial Pneumonia (c) Viral Pneumonia (d) Normal.

### III. PROPOSED MODEL

Presently, the most popular machine learning method is the convolution neural network (CNN), a kind of deep artificial neural network with superior feature extraction and identification capabilities [14, 15]. One of its most distinguishing characteristics is weight sharing, which drastically decreases the network complexity and total weights [16]. With numerous convolution layer, followed by pooling layer, activation functions like softmax, ReLu, and fully connected layers reduce the loss function to a minimum.

The ultimate result of a typical CNN network structure is often one or more fully-connected layers [17, 18, 19], however the network itself is typically composed of convolutional and pooling layers. The final output of the layer is obtained by applying a bias to fully connected layer and activation function.

$$D_j^l = f\left(\sum_{D \in M_j} D_j^{l-1} \cdot h_{ij}^l + v_i^l\right) \quad (1)$$

The Eq. 1 illustrates this process. feature map is represented as  $D_j^l$ , excitation function is shown as  $f$ , input feature maps is shown as  $M$ , convolution operation is mentioned as  $\cdot$ , and bias term is represented as  $h$ .

### IV. DEEP RESIDUAL NETWORKS

ResNet is built on top of deeper networks, the theoretical foundation of which is described in [20, 21]. There are 50, 101, and 152 nodes in a traditional ResNet network. The 2015-ILSVRC competition was won by a CNN with 152 layers. In addition, ResNet improves by 28% on the well-known cOco132 example dataset for image recognition [22, 23]. For the most part, ResNet takes use of the concept of bypass channels in the "road network," as seen in the mathematical Eq. 2 and Eq. 3.

$$g(y_i) = f(y_i) + y_i \quad (2)$$

$$f(y_i) = g(y_i) - y_i \quad (3)$$

The transformed signal, denoted by  $f$  in Eq. 2 and Eq. 3, is defined in relation to the input signal,  $y_i$ . A side-channel adds the first input to  $f(y_i)$ . When computing the residual in Eq. 4,  $g(y_i)$  is employed. In order to facilitate communication across layers, ResNet deploys "shortcut channels" inside those levels; in contrast to the gates used in traditional highway networks, however, these gates are both data-agnostic and parameter-free. They stand for the non-residual functions of a highway system after the bypass route has been closed. A linear direct mapping



$y \rightarrow y_i$  and a nonlinear mapping  $F$  are both possible components of the submodule  $f(y_i)$ . Once the learning algorithm determines that the direct mapping,  $y \rightarrow y_i$ , is best, it may simply zero out the weight parameters of the nonlinear mapping  $f(y_i)$ . When there is no one-to-one mapping, it might be challenging to learn a linear mapping from a nonlinear function  $f(y_i)$  [24, 25]. On the other hand, with ResNet, residual information is continuously sent and the shortcut channel is never closed. In order for ResNet to circumvent the gradient reduction issue, it makes use of residual links, which are rapid channel connections that speed up the fusing of deep networks. The architecture of ResNet50 is shown in Fig. 3.

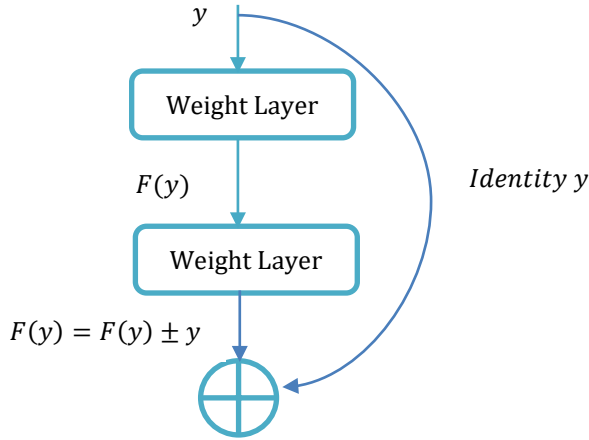


Fig. 3. ResNet Network Architecture.

**Algorithm:**

Input: Feature vector from CNN model( $F$ ), Size of  $F(S)$ ,  $F$  average (Avg),  $F$  Standard Deviation (SD), Threshold ( $T$ ), Parameter Tuning ( $F(X)$ )

Output: Reduced Feature ( $OF$ )

Step 1: *Feature Reduction* ( $F, SD, Avg, T$ )

Step 2: Begin

Step 3:  $OF = F$

Step 4: *For*  $i = 1$  to  $S$  *do*

Step 5:  $X = SD/OF[i]$

Step 6:  $Y = Avg/OF[i]$

Step 7: *if*  $X > T$  and  $Y > T$

Step 8: Parameter Tuning

$$F(X) = F(X) \pm X \text{ and } F(Y) = F(Y) \pm Y$$

---

Step 9:	$OF[i] = []$
Step 10: end if	
Step 11: end for	
Step 12: end	

---

V. RESULTS AND DISCUSSIONS

We started by analysing X-ray and CT scans of the lungs taken from a wide range of patients and healthy individuals (<https://www.kaggle.com/datasets/tawsifurrahman/covid19-radiography-database>). We gathered X-ray and CT scans of the lungs from 2045 healthy controls, 785 mild cases, and 130 severe cases of COVID-19. Data sets were obtained and manually annotated by expert imaging specialists, and pictures were utilised with patient and hospital permission. Ethical clearance is given by the institution's review board. A total of 2960 samples were obtained. All pictures were exactly 299 pixels in width and height. Table II displays the experimental data distribution. We combine COVID-19 features with clinical knowledge to segment the lesion area, to train a neural network to predict COVID-19 diagnoses, ultimately narrowing down the features to 20 that have the highest diagnostic value. This offers a non-invasive technique for detecting COVID-19 in advance.

Chest X-Ray images from Kaggle: 6110 pre-processed images are used for training and testing. If model can train in a shorter time, this means that model will be more efficient at more training iterations. As it can be seen on the Table I below DenseNet-121 model is more in the aspect of time. It is common knowledge that classification algorithms are evaluated based on parameters like specificity, sensitivity, and accuracy as shown in Eq. (4), Eq. (5), and Eq. (6). And proposed model attain values are mentioned in Table II. Accuracy is defined as the ratio of correctly categorized instances to the total number of examples. The output images are represented in Table III. Proposed method compared with the existing method are shown in Table IV.

Accuracy is compared with number epoch as shown in Fig. 3.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{4}$$

$$Sensitivity = \frac{TP}{TP+FN} \tag{5}$$

$$Specificity = \frac{TN}{TN+FP} \tag{6}$$

TABLE I. DEEP LEARNING BASED CLASSIFICATION

Year	Author	Model	Dataset	Application	Remarks
<b>Supervised classification</b>					
2019	Schlemper et al.	AG-Sononet: attention-gated model	ChestX-ray14	image plane classification using 2D fetal ultrasound	To better leverage local information and aggregate attention vectors at various sizes for final prediction, Baumgartner et al. (2017) integrated grid attention.
2018	Guan et al.	AG-CNN: an attention guided CNN	Private dataset	chest X-rays Thorax disease classification	The global picture was parsed for discriminative areas using an attention mechanism, and those regions were utilized to train a local CNN node
<b>Unsupervised image synthesis</b>					
2018	Wu et al.	cGAN	DDSM dataset	Mammogram classification	Using labels of malignant and non-malignant to exercise control over the development of a certain kind of lesion.
2018	Frid-Adar et al.	ACGAN	Private dataset	CT liver lesion classification	Analyzing the differences in performance of GAN's
<b>Self-supervised learning based classification</b>					
2019	Chen et al.,	Common CNN Structure	Private dataset	Fetal ultrasound image plane classification	Developing a novel context-restored self-supervised method
2021	Azizi et al.	MICLe: based on SimCLR	CheXpert, and private dataset	Chest X-Ray Image classification	Proposing a novel contrastive learning strategy based on SimCLR that uses several pictures for self-supervised pre-training.
2021	Vu et al.	MedAug: Based on MoCo	CheXpert	Pleural effusion based chest X-ray classification	Self-supervised pre-training outperforms ImageNet pre-training.
2021	Zhou et al	Models Genesis	LUNA 2016	CT lung nodule for false positive reduction.	Combining four self-supervised techniques to learn from diverse viewpoints (appearance, texture, and context)

TABLE II. PROPOSED MODEL TRAINING TIME AND ACCURACY RATE

Model	Training Time	Accuracy Rate	Accuracy on Testing Data
ResNet	512 secs	88.7	4/100
DenseNet	248 secs	84.9	6/100
Featured ResNet	214 secs	88.2	6/100

TABLE III. OUTPUT IMAGES CLASSIFICATION


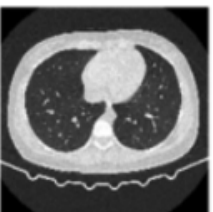
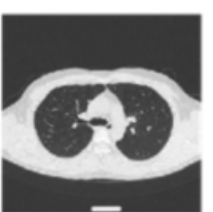


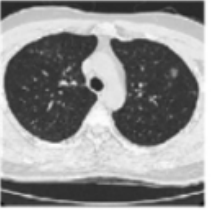

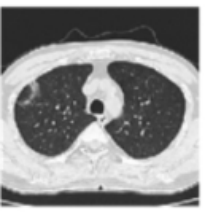
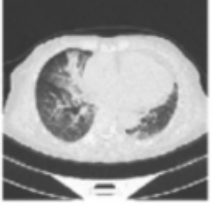
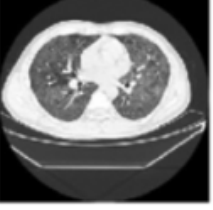

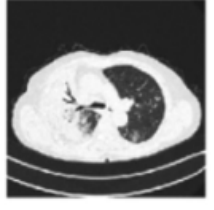
<b>Normal</b>				
<b>Mild</b>				
<b>Severe</b>				

TABLE IV. PROPOSED CNN VS. EXISTING CNNs

Deep Models	Accuracy	Sensitivity	Specificity	TP	FP	FN	TN	MCC	F Score
Squeeze Net	97.874	0.9	0.9	609	19	20	618	0.92	0.95
DenseNet201	97.564	0.9	0.89	610	25	18	612	0.92	0.96
Xception	96.644	0.89	0.89	624	29	28	634	0.92	0.94
Inceptionv3	98.044	0.91	0.91	610	24	26	612	0.92	0.94
Google Net	97.344	0.9	0.89	609	26	20	614	0.92	0.95
Resnet50	97.574	0.89	0.89	613	22	21	614	0.93	0.94

## VI. CONCLUSION

The approach described in this work has been developed primarily for the purpose of making early and definitive diagnoses for patients, although it has broad use in pathological categorization and prognosis prediction. When compared to the riskier and slower nasal swab, CNN is a preferable approach for identifying people infected with COVID-19. Importantly, the proposed approach allows for a multiclass diagnosis with other pulmonary disorders, which is important since many of these diseases may have startling similarities in their symptoms and consequences on the lungs. The superiority of CNN models like COVDC-Net in real-world applications may soon become apparent, rendering other testing obsolete. Another advantage of AI-assisted diagnosis is that it may be quickly scaled to existing hospitals and clinics using X-ray machines, without the need for specialised infrastructure and testing equipment. We want to improve the suggested model's robustness in the future by testing it on a more comprehensive range of pulmonary disorders.

## REFERENCES

- [1] F He, Y Deng, W. Li, Coronavirus disease 2019: What we know? *J. Med. Virol.* 92 (7) (2020) 719–725.
- [2] Felsenstein, J A Herbert, P S McNamara, et al., COVID-19: Immunology and treatment options, *Clin. Immunol.* 215 (2020) 108448.
- [3] J. Li, L. Liu, S. Fong, R.K. Wong, S. Mohammed, J. Fiaidhi, Y. Sung, K.K.L. Wong, Adaptive Swarm Balancing Algorithms for rare-event prediction in imbalanced healthcare data, *PLoS One* (2017), doi:10.1371/journal.pone.0180830.
- [4] Y. Ye, J. Shi, Y. Huang, D. Zhu, L. Su, J. Huang, Management of medical and health big data based on integrated learning-based health care system: a re- view and comparative analysis, *Comput. Method. Program. Biomed.* 209 (2021) 106293.
- [5] C Stasi, S Fallani, F Voller, C. Silvestri, Treatment for COVID-19: an overview, *Eur. J. Pharmacol.* 889 (2020) 173644 Epub 2020 Oct 11. PMID: 33053381; PM- CID: PMC7548059, doi:10.1016/j.ejphar.2020.173644.
- [6] K. R. Devi, S. Suganyadevi, S. Karthik and N. Ilayaraja, "Securing Medical Big data through Blockchain technology," 2022 8th International Conference on Advanced Computing and Communication Systems (ICACCS), 2022, pp. 1602-1607, doi: 10.1109/ICACCS54159.2022.9785125.
- [7] Suganyadevi, S., Seethalakshmi, V. CVD-HNet: Classifying Pneumonia and COVID-19 in Chest X-ray Images Using Deep Network. *Wireless Pers Commun* (2022). <https://doi.org/10.1007/s11277-022-09864-y>.
- [8] S. Suganyadevi, K. Renukadevi, K. Balasamy and P. Jeevitha, "Diabetic Retinopathy Detection Using Deep Learning Methods," 2022 First International Conference on Electrical, Electronics, Information and Communication Technologies (ICEEICT), 2022, pp. 1-6. <https://doi.org/10.1109/ICEEICT53079.2022.9768544>.
- [9] Suganyadevi, S., Seethalakshmi, V. & Balasamy, K. A review on deep learning in medical image analysis. *Int J Multimed Info Retr* (2021). <https://doi.org/10.1007/s13735-021-00218-1>.
- [10] Balasamy K, Suganyadevi S (2021) "A fuzzy based ROI selection for encryption and watermarking in medical image using DWT and SVD" *Multimed Tools Appl* 80, 7167–7186, <https://doi.org/10.1007/s11042-020-09981-5>.
- [11] R Citro, G Pontone, M Bellino, et al., Role of multimodality imaging in evaluation of cardiovascular involvement in COVID-19, *Trend. Cardiovasc. Med.* 31 (1) (2021) 8–16.
- [12] M J Horry, S Chakraborty, M Paul, et al., COVID-19 detection through transfer learning using multimodal imaging data, *IEEE Access* 8 (2020) 149808–149824.
- [13] Jianshe Shi, Yuguang Ye, Daxin Zhu, Lianta Su, Yifeng Huang, Jianlong Huang, Comparative analysis of pulmonary nodules segmentation using multiscale residual U-Net and fuzzy C-means clustering, *Comput. Method. Program. Biomed.* 209 (2021) 106332.
- [14] T J Brinker, A Hekler, J S Utikal, et al., Skin cancer classification using convolutional neural networks: systematic review, *J. Med. Internet Res.* 20 (10) (2018) e11936.
- [15] R. Girshick, Fast r-cnn, in: Proceedings of the IEEE international conference on computer vision, 2015, pp. 1440–1448.
- [16] H C Shin, H R Roth, M Gao, et al., Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning, *IEEE Trans. Med. Imaging* 35 (5) (2016) 1285–1298.
- [17] Balasamy, K., Krishnaraj, N. & Vijayalakshmi, K. Improving the security of medical image through neuro-fuzzy based ROI selection for reliable transmission. *Multimed Tools Appl* 81, 14321–14337 (2022). <https://doi.org/10.1007/s11042-022-12367-4>.
- [18] S Hershey, S Chaudhuri, D P W Ellis, et al., CNN architectures for large-scale audio classification, in: 2017 IEEE international conference on acoustics, speech and signal processing (icassp), IEEE, 2017, pp. 131–135.
- [19] Gopalakrishnan T., Ramakrishnan S., Balasamy K., Murugavel A.S.M., Semi fragile watermarking using Gaussian mixture model for malicious image attacks, 2011 World Congress on Information and Communication Technologies, 2011: 120 – 125.
- [20] L Yu, B Li, B. Jiao, Research and implementation of CNN based on TensorFlow, *IOP Conference Series: Materials Science and Engineering*, 490, IOP Publishing, 2019.
- [21] Z Lu, Y Bai, Y Chen, et al., The classification of gliomas based on a pyramid dilated convolution resnet model, *Pattern Recognit. Lett.* 133 (2020) 173–179.
- [22] Krishnasamy, B., Balakrishnan, M., Christopher, A. (2021). A Genetic Algorithm Based Medical Image Watermarking for Improving Robustness and Fidelity in Wavelet Domain. In: Satapathy, S., Zhang, YD., Bhateja, V., Majhi, R. (eds) *Intelligent Data Engineering and Analytics. Advances in Intelligent Systems and Computing*, vol 1177. Springer, Singapore. [https://doi.org/10.1007/978-981-15-5679-1\\_27](https://doi.org/10.1007/978-981-15-5679-1_27).
- [23] Suganyadevi S, Shamia D, Balasamy K (2021) An IoT-based diet monitoring healthcare system, for women. *Smart Healthc Syst Des Secur Priv Asp.* <https://doi.org/10.1002/9781119792253.ch8>.
- [24] L Wen, X Li, L. Gao, A transfer convolutional neural network for fault diagnosis based on ResNet-50, *Neur. Comput. Applica.* 32 (10) (2020) 6111–6124.
- [25] K He, X Zhang, S Ren, et al., Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

# Gamification on OTT Platforms: A Behavioural Study for User Engagement

Komal Suryavanshi<sup>1</sup>, Prasun Gahlot<sup>2</sup>, Surya Bahadur Thapa<sup>3</sup>, Dr. Aradhana Gandhi<sup>4</sup>, Dr. Ramakrishnan Raman<sup>5</sup>

Research Scholar, Symbiosis Center for Research & Innovation (SCRI), Symbiosis Centre for Behavioural Studies, Symbiosis Institute of Business Management, Symbiosis International (Deemed University), Pune, India<sup>1,2,3</sup>

Professor (Retail and Marketing), Symbiosis Centre for Behavioural Studies, Symbiosis Institute of Business Management, Symbiosis International (Deemed University), Pune, India<sup>4</sup>

Director, Symbiosis Institute of Business Management, Symbiosis International (Deemed University), Pune, India<sup>5</sup>

**Abstract**—This study examines the consumer's visual attention toward gamification options while watching the OTT (Over-the-top) online content. Also, the impact of gamification on user engagement (UE) on the OTT platform was studied using data collected by conducting an eye-tracking experiment and subsequently using a user engagement scale (UES). The study was carried out at the marketing and behavioural lab of a management institute in India using the OTT platform website and Tobii eye-tracker. Empirical data was collected from 52 respondents within the age group between 23 to 35 years. The relation between Attention to Gamification (AG), Reward Satisfaction (RS), and User Engagement (UE) were studied by running a mediating linear regression analysis. From the results, it was found that respondents were equally interested in watching the online content as well as ready to explore the gamification options. The research findings demonstrate that Reward Satisfaction (RS) acted as a mediating factor in the relation between Attention to Gamification (AG) and User Engagement (UE). This study adds to the literature on consumer engagement towards gamification on the OTT platform, where the literature is still limited. Future research could consider mobile apps as a platform to undertake the study. This study aimed to empirically test the effect of AG on UE with the involvement of RS as a mediator. The study is the first of its type to use eye-tracking data to understand the impact of gamification on the OTT platform.

**Keywords**—Gamification; user engagement; eye-tracking; OTT (Over-the-top); reward; visual attention

## I. INTRODUCTION

Technology is proliferating, and the internet offers a new way of entertainment, which has augmented the adoption of Over-the-top (OTT) media. The OTT platforms are a significant by-product of the expansion and exploration of digital media. The OTT market is anticipated to grow at a compounded annual growth rate (CAGR) of 14.3 % to reach US \$86.80 billion by 2026 [1]. This growth is due to easily accessible and inexpensive internet connections and low-cost subscription-based OTT platforms. With technological advancements in handheld devices, viewers can see content, including videos, movies, series, etc., anytime and anywhere. OTT service platforms like Netflix, Disney Hotstar, Amazon Prime, Discovery Plus, YouTube, EPIC ON, ZEE5, and others enable viewing movies, series, TV shows, and other content at a click. The OTT market has steadily and slowly shaken the

linear, complex, and vertically unified television distribution industry, which has been dominated for years by traditional pay television channels [2]. The stagnancy of the conventional television market created a vacuum in the media market. OTT platforms have seized this chance to strategize about how to engage with their customers more effectively, micro-target them, customize their products, and take advantage of the impending media collapse. This has eventually created an inspiring and level-playing scenario for OTT content providers [3]. Thus, to stay in the competition, many OTT platforms have implemented gamification to provide value-added engagement to customers [4]. Gamification is implemented to increase awareness about the gamified platform, achieve a prominent presence in the market, develop its user base, strengthen the bonds with the viewers, and improve consumer engagement.

With multiple players in the OTT market, operators aim to understand and offer services that cater to the viewer's demand by analysing the subscriber's preferences, thereby transforming how consumers consume online media content. The streaming platforms provide viewers with the freedom to select and access the content of their choice. Many platform providers have adopted operational changes concerning content creation, representation, and delivery to attract customers and gain their loyalty, as well as the introduction of gamification to encourage customers to participate actively. Games are essential to motivation and engagement when applied to interactive platforms [5]. Gamification refers to playing games on the platform to achieve rewards that can be redeemed in return and generate consumer engagement [6]. It is a concept of improving services with value addition to gaming experiences [7]. With gamification, consumers play games and enjoy the experience, irrespective of the result, thus increasing their involvement with the gamified platform [8] [9]. With gamification tools, consumers participate in interactive games, puzzles, questions, fantasy points, leader boards, bids, auctions, badges, feedback challenges, performance or game progress bars, lotteries, countdowns, and other options alongside viewing content. Affordances (points, badges, and leader boards) in gamification are the elements that form game structure and induce gameful experiences. The user interfaces with gamification and rewards (incentive to play games) shape consumer behaviour [8]. Gamification has attracted many retailers and e-commerce giants to include it for engaging customers with loyalty programs, impacting consumers' buying decisions and incremental sales [10]. The point-of-purchase

marketing was improved for offline and online stores that included gamification in the purchase process [11]. Gamification impacts viewers' behavioral outcomes, increasing platform involvement, and engagement [7]. Gamification in the marketing campaign comprises four levels: "attract, engage, retain, and reward" [12], with the ultimate goal of enhancing participation. Investigating gamification implementation in the context of the OTT platform is of significant importance as the OTT platforms are featured with online content viewing, which leads to involvement, interaction, and playful machination of the everyday world. Gamification enhances motivation [13]; therefore, the impact of gamification on customers' intention to engage on the OTT platform deserves investigation.

Concerning the OTT platforms, studies have mainly concentrated on examining the factors impacting the adaption and adoption of OTT platforms by the viewers [14]. Other studies include the OTT business models and approaches for business extension [15]. Even though OTT platforms have become part of everyday media, the studies related to the adoption and effect of gamification on OTT platforms are limited. ZEE5 (OTT platform) started ZEE5 Super Family (ZSF), a gaming experience for fictional content viewers [4]. EPIC ON has incorporated games that let the viewers redeem the rewards with coupons and discounts. Thus, this study tries to understand if implementing gamification on the OTT platforms improves customer engagement with the following research questions-

RQ1: Does the customer pay equal attention to the gamification option while watching the online content on the OTT platform?

RQ2: Does the option to earn rewards on playing the game act as a mediator between attention to gamification and customer engagement on the OTT platform?

To address the above research questions, this study considered the EPIC ON OTT website platform and collected users' visual attention data from an eye-tracking device to investigate the effectiveness of gamification on OTT. Indian OTT market is set to arise as the next biggest OTT market to reach the value of ₹ 138 billion by the end of the financial year 2023 with an estimated growth of 45%, following the USA [3]. OTT platforms have therefore been forced to implement structural changes to improve user engagement and retention. Many OTT service providers have embedded gamification into their platforms. This technological and structural update provides an important context that answers the research questions.

The objective of this study is two-fold. Firstly, a lab experiment was conducted using eye-tracking software to investigate customers' visual attention towards the option to watch and play the game on the OTT platform. Secondly, the study examined the impact of reward as a mediator for attention to gamification and user engagement. With the growing competitive environment in the local and international media marketplaces produced by global OTT services, it is vital to develop a new strategy for each OTT service through research that considers the viability of novel techniques like gamification. As of yet, the OTT platform is used to view content, but playing games on the OTT platform while

watching content is a new concept. The results from this study contribute to the body of literature on OTT-based gamification and provide developers an insight into whether the implementation of gamification on OTT is feasible.

The remainder of this research paper is comprised as follows. First, the report starts with the literature review and formulates the hypotheses. Then, the paper presents the methods and outcome of the analysis. Thirdly, arguments on the findings and implications are mentioned. Lastly, the report provides limitations and directions for future research.

## II. LITERATURE REVIEW

### A. Gamification Research

Extensive use of the term *gamification* started an era ago [16]. Since its commencement, it has been implemented in diverse fields, including computer sciences, educational scenarios, the health care sector, tourism, governance, research, and marketing [17][18][19], to name a few. Gamification means utilizing game elements and collaborating with various platforms to improve engagement. It is an approach to implementing game design components in the non-gaming environment [16]. Gamification has been defined as "the process of using game mechanics with other forms of technology to increase engagement" [20]. Most of the description of gamification that has been published state that it adopts game-like strategies in non-game contexts and can engage users and produce value that users perceive [21]. When implementing gamification, the creation of customer engagement is necessary. If the users do not experience participation, the whole gamification process fails [7][6][16].

A few quantitative research studies have shown the causal relationship between gamification, purchasing decisions, and customer engagement. In human-computer interaction, "user engagement" describes how people interrelate with the technology that fascinates them. A study by [22] stated that gamification improves adoption via playfulness, making consumers curious about the innovative features and their relative advantage. With gamification, consumers comment on products or services, give reviews, and share the content, increasing active users and repeated visits [23]. Gamification consists of stages like- appeal, involve, hold, and monetize [12], with the final goal to engage customers on the platform [24]. Involvement in technology results in engagement from the interaction between an emotional, cognitive, and behavioural relationship [25]. Gamification has been successfully applied across the learning management system for the students, motivating and generating interest in the subjects [26]. Employee engagement and job interest have been seen to be improved with gamification [9]. Thus, much of the research in the past has focused on understanding the role of gamification in encouraging customer participation and enhancing engagement in areas including social media, e-commerce, fitness apps, etc. Previous researchers have considered engagement and its effect on users; these ideas involve studies related to loyalty [27][28], pleasure [27], conviction [29], commitment and emotional connection [30]. However, gamification has been implemented recently in the case of OTT platforms, so the research conducted to understand the influence on viewer engagement is limited [31].

Providing options to play games on the OTT platform offers gameful experiences for the users regardless of the outcomes [7]. The present work attempts to study the gap using an experimental design in a lab setting with an eye-tracking tool.

### B. OTT Platform

The increased internet penetration and availability of multiple media platforms have stimulated video consumption via digital platforms. OTT service platforms, adopted from the TV set-top box, distribute video content using internet protocol. Reference [32] indicated that there had been a growing tendency toward the consumption of OTT platform content compared to traditional TV. Bypassing cable and satellite transmission, OTT video streaming services are defined as digital platforms providing consumers with handpicked content worldwide. A study by [33] stated that the pattern of complete streaming seasons for instant consumption has acted as a seed for the change from the television viewing culture to OTT content viewing. With the rapid growth in the media market, research studies have focused on user acceptance behaviour towards OTT platforms. Studies show that the content streamed over the platform and the involvement of consumers with it has played an essential role in consumers' reception and loyalty to the forum. This has led to unique content on the OTT platform that ensures enhanced experiences and consumer engagement [1]. Previous studies on consumer behaviour towards OTT platforms adopted expectation confirmation theory (ECT) [34] to understand the continued use intention of the consumers. The technology acceptance model (TAM) [35] was used to understand the motivation systems theory [36] regarding the consumer's choice of OTT platforms. With the increase in the number of OTT services and the number of viewers, [37] carried out user-centric research to study the user experience on the OTT platform (Netflix) based on uses and gratification theory

(UGT) and TAM. Reference [38] used niche analysis and the Uses and Gratification Theory (UGT) to inspect the viewers' interest in the OTT platforms. As OTT platforms are in their innovative stage, the service providers are experimenting with different tools to increase engagement. Reference [39] conducted a literature review study on 262 articles based on the OTT phenomenon, adopting experimental analysis, descriptive analysis, case study method, survey, content analysis, and theoretical analysis. However, in India, the majority of the studies on the OTT have used survey techniques. Thus, this study fills a research gap by adopting a novel methodology like experimental design using eye tracking in the Indian context.

### C. Customer Engagement (CE) Research

Customer engagement (CE) research has gained momentum. Gamification analysis primarily displays platform engagement [40]. Attention has been defined differently across various academic disciplines [41], and numerous definitions have been used to describe diverse engagement objects and subjects (e.g., brand engagement, customer engagement, student engagement, user engagement, employee engagement). In recent years, multiple studies have researched the relationship between gamification and various forms of engagement. Gamification implementation to motivate and engage students in academics has received the most attention, with education being the most fertile research field [42]. Nevertheless, research concerning engagement and gamification in contexts apart from education is getting attention rapidly. As mentioned in Table I, research studies have investigated the link between gamification and brand engagement [40],[60],[49]; customer engagement (e.g., [24], [9], [47], [48]); employee engagement (e.g., [61]) and user engagement (e.g. [55], [57]). The current study focuses on user engagement, which is driven by gamification and related reward.

TABLE I. STUDIES INVESTIGATING THE RELATIONSHIP BETWEEN ENGAGEMENT AND GAMIFICATION

Reference	Independent variables	Mediator/Moderator	Dependent variables	Research design	Key findings
<b>Customer engagement</b>					
Reference [24]	Game elements (challenge, tasks, rewards, badges, leader boards and win condition)	Customer engagement behaviours and customer engagement emotions	Reward, relationship, loyalty and Subversion	Geographic approach	The study identifies essential behaviours and processes of online customer interaction.
Reference [43]	Gamification mechanics for player types		Customer and employee Engagement	Case Study	Gamification may increase employee and customer engagement, enhancing how people interact with brands and businesses and boosting workplace efficiency.
Reference [44]	Gamification mechanics	Challenge, entertainment, social dynamics and escapism/ Medical predispositions and age	Patient engagement (cognitive, emotional and behavioural)	Case study	Patient engagement is increased due to the four experiential outcomes that gamification mechanisms generate in patients: challenge, amusement, social dynamics, and escape.
Reference [45]	Perceived usefulness, ease of use, social influence and	Customers' engagement intention	Brand attitude	Focus group and survey	Perceived enjoyment and usefulness forecast brand attitude and engagement intentions. These characteristics are not influenced by perceived ease of use. Only brand attitude is influenced by



	enjoyment				perceived social impact.
Reference [46]	Game elements		Brand awareness, tourist experiences, tourist engagement, customer loyalty, entertainment and employee management	Case study	Gamification can be used in tourism marketing
Reference [47]	Gamified customer benefits (epistemic, social integrative and personal integrative)	Age and experience	Customer engagement behaviour and purchase	Longitudinal design	Personal and social consolidative reimbursements are the best drivers of engagement and purchase
Reference [44]	Game elements (competition and cooperation)	Customer experience, losing a contest/Prior level of customer engagement	Customer engagement toward the co-creation activity (conscious attention, enthused participation and social connection) and community	Experiment	Win/lose choices fall apart the benefits of gamification. Losing a competition incorporates an adverse effect on client encounters and engagement.
Reference [48]	Gamification principles (social interaction, sense of control, goals, progress tracking, rewards and prompts)	Hope, compulsion, customer engagement	Purchases	Interviews and survey	Trust emphatically intercedes the relationship between gamification standards and client engagement. Compulsion decreases the plausibility of client engagement.
<b>Brand Engagement</b>					
Reference [49]	High interactivity; optimal challenge	Emotional brand engagement; cognitive brand engagement	Self-brand connection	Experiment	Gamified communications that are highly collaborating and optimally challenging enable self-brand connections.
Reference [24]	Challenge, tasks, rewards, badges, leader board, and win condition	Customer engagement behaviours, fun/enjoyment (flow), dissatisfaction	Reward, relationship, loyalty, subversion	Geographic approach	The findings distinguish primary forms and results of C and CEB inside virtual gamified stages.
Reference [50] [51]	Perceived mobility, utilitarian and hedonic features	User experience; perceived benefits; perceived values	Brand equity (perceived quality, loyalty, associations, trust)	Web-based survey	Mobility has a significant effect on functional & hedonic features, while mobility and utilitarian and hedonic features influence consumer experience, which affects brand fairness.
Reference [18]	Gamification mechanisms		Brand engagement, brand loyalty, and brand awareness	Case study and interviews	Marketing executives see increased engagement as one of the most vital benefits of gamification.
Reference [40]	Gamification		Consumer brand engagement and consumer benefits (functional, hedonic, social, and educational)	Interviews	Gamified packaging generates: hedonic, functional, social, and academic edges for the client that are coupled to consumer whole engagement dimensions (cognitive, emotional, and behavioural)
Reference [52]	Immersion-, achievement- and social-related gamification features	Brand engagement (cognitive, emotional and behavioural)	Brand awareness and brand loyalty	Survey	Achievement and social interplay-related gamification feature positively impact the three varieties of company engagement. Immersion-related gamification aspects are solely positively associated with social brand engagement. Brand engagement will increase company consciousness and brand loyalty.
<b>User Engagement</b>					
Reference [53]	Game Design Mechanisms		Engagement with online platforms (Objective metrics)	Experiment	Gamified thematic activities, graphical incentives, and discussion boards influence member retention and engagement.
Reference [54]	Game elements (points, rankings, achievement and		Engagement toward a computation system, acceptance (attitude, intention to use, and intention to	Experiment	Respondents experience added engagement and show higher behavioural intents toward the gamified system. Perceived output

	social elements)		recommend), perceived usability and perceived output quality		quality and perceived engagement influence the reception of the gamified system
Reference [55]	Game dynamics (rewards, competition, self-expression, altruism)	Competence, autonomy, relatedness and enjoyment	User engagement with a gamified information system (vigor, dedication, and absorption)	Survey	Gamification improves user engagement by mediating psychological needs, satisfaction (autonomy, competence, and relatedness), and fun.
Reference [56]	Game elements (score system, a progress bar and levels, leader board, feedback)		TAM (perceived utility, ease of use, external factors, attitude towards and demonstrated results) and user engagement with a health mobile app (focus and attention, usability perception, aesthetic aspects, supportability, originality, and involvement)	Experiment	Gamification impacts engagement positively, inspiring intrinsic motivation in the respondents.
Reference [57]	Game elements (competition and leader boards)		Engagement with an app (objective metric)	Experiment	Gamification intensifies engagement with the app
Reference [58]	Commensurate game elements (e.g., points) and incommensurate elements (e.g., likes)		Autonomy, competence, relatedness, engagement behaviour (objective metrics), intrinsic motivation, loyalty	Experiment	Users who engage with equivalent game features exhibit greater internal motivation, are more involved in physical activity, and are more devoted to the fitness app than users who do not.
Reference [59]	Perceived usefulness, perceived ease of use, convenience and enjoyment	Engagement with mobile apps	Intention to use	Survey	Perceived ease of use, perceived usefulness, and enjoyment influence engagement, leading to users' intention
Reference [13]	Gamification Design (Badges)	Disparity in professional seniority	Engagement with online platforms (objective metrics) and inequality economic of returns	Experiment	Gamification design boosts doctors' participation in online health communities.

The User Engagement Scale (UES), developed by [62], is the most popular measure for user engagement. The original UES included 31 components across six user engagement dimensions (i.e., aesthetic appeal, felt involvement, novelty, perceived usability, focused attention, and durability). The factors included the following- Focused Involvement (FI) (if the experience is enjoyable or intriguing); Focused Attention (FA) (focused concentration, absorption, and the loss of the sense of time); Endurability "EN" (holistic response to the experience and overall success of the interaction); Novelty "NO" (interest or curiosity generated by the system throughout the buying task); Perceived Usability ("PU") and Aesthetic Appeal ("AE") are two terms for the visual look of an interface, which includes the visuals, graphics, and other items that appeal to the user's senses (affective and cognitive aspects derived from the use of the system).

Understanding individuals' communicative patterns with digital platforms (e.g., eHealth, eLearning, digital games, social media, online search) is essential in studying their effects on user behaviour [63]. With a wide variety of digital platforms (e.g., social networking sites, mobile apps, web search engines, and others), the association between gamification and user engagement has been investigated in various contexts like online platforms, mobile apps, learning management systems, human computation and others (Table I). These studies

established a positive relationship between gamification and users' engagement.

### III. PROPOSED MODEL AND HYPOTHESES: GAMIFICATION AND ENGAGEMENT

#### A. Study 1

For the first part of the study, the data from the eye-tracking device was used to examine the users' visual attention towards the option to watch online content and the opportunity to play games. Eye-tracking data analysis helps to find patterns in the respondent's visual data. However, finding those patterns and analyzing the eye-tracking data require more than one visualization metric. Using multiple metrics to analyze the eye-tracking data can improve the result [64]. Fixation count (location), fixation length, and saccades (movement) are the essential pieces of eye-tracking data that form important parameters [65]. When the eye is reasonably still, it is called a fixation. Fixations determine where participants fixate their eye vision when viewing the platform. Visualization data regarding fixation duration and the count was used for the present study. Fixation Duration determines how long people spend staring at a specific spot. Each interval is a few milliseconds long. Processing is linked to increased duration, indicating complexity, interest, or engagement. Fixation counts determine which portion of the page receives more or less attention. Fixation counts are numbered in sequence to see how people

process survey pieces. Studies based on advertisements [66] included visual attention data to analyze the impact on viewers. Therefore, based on the above statements, the following hypotheses are proposed.

H1: Duration of visual attention (fixation duration) towards watching online content and gamification is the same.

H2: Visual attention, using fixation counts across watching online content, and gamification are the same.

### B. Study 2

For the second part of the study, a model in Fig. 1 is proposed to understand if the viewer's attention to gamification impacts engagement with reward as a mediator. The data based on the user's visual attention on a platform displays the depth of user involvement on the platform and is an aspect of the user experience [67]. Reference [23] and [68] analyzed the visual attention data to understand user engagement. Gamification elements impact the perceived ease of use of shopping and e-banking websites [69]. Many gamified services introduce challenges as one of the prime game elements [70]. Previous research studies have shown that challenges in gamified shopping impact buying behaviour [71]. A reward gained by completing a task increases the likelihood of acting on that reward [11]. This effect has been explored in situations other than gamification and intrigues the effort necessary to overcome the challenge. Earning a reward improves the likeliness of the tip compared to receiving a prize by luck [72]. Benefits such as coupons, discounts, cashback offers, or free subscriptions motivate and engage users. A reward is considered an essential aspect of engaging customers in gamification, and thus, the following conceptual model is proposed, as given in Fig. 1.

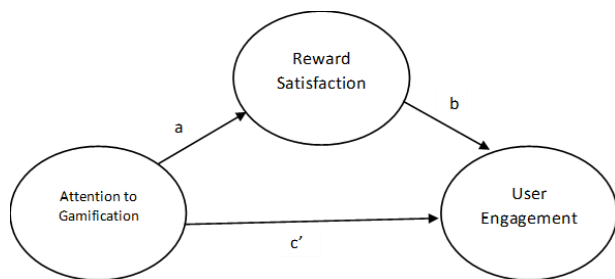


Fig. 1. Attention to Gamification and its Impact on user Engagement with Reward Satisfaction Acting as a Mediator.

1) *Attention to gamification and user engagement:* In their study, [73] stated that the adoption of gamification significantly leads to customer engagement. Reference [9] cited gamification as a game design scheme to achieve customer engagement and retention. Similarly, organizations use gamification to motivate and engage employees and customers [74]. Reference [48] suggested that gamification helps firms achieve customer engagement via social interactions. Reference [51] supported using gamification in marketing activities to strengthen customer engagement. Reference [75] stated that gamification's adoption boosted stakeholder engagement and is the primary driver of customer engagement. Tourist destinations employ gamified journeys to

engage visitors through various elements such as interactive maps, challenges, rewards gain, plot, and other aspects [76]. Thus, to test the impact of gamification on user engagement when implemented on OTT, the following hypothesis is formulated:

H3: There is a positive relationship between Attention to Gamification on the OTT platform and User Engagement.

2) *Reward satisfaction and user engagement:* Gamification on a gamified platform involves dealing with obstacles or being challenged [77], and dealing with such contests takes a certain amount of engagement. Continued engagement intent is driven by reward fulfillment. An extrinsic motivator, such as a reward, is a particular outcome of an activity that seeks to influence a person toward a specific action (like playing games). A reward represents a distinct outcome from the act itself, making it an extrinsic incentive and a motivator [78]. In the gamification literature, rewards are frequently used to incentivize activity. Users feel their connections with the gamified system are valuable and meaningful if they achieve rewards via participation [11]. In our study, on playing games, users were rewarded in the form of coupons, discounts, and redemption in monetary or non-monetary conditions. As rewards act as an extrinsic motivator to improve user engagement, the following hypotheses is proposed:

H4: There is a positive relationship between Attention to Gamification (AG) and Reward Satisfaction (RS).

H5: There is a positive relationship between Reward Satisfaction (RS) and User Engagement (UE).

H6: Reward Satisfaction (RS) mediates the effects of Attention to Gamification (AG) and User Engagement (UE).

## IV. METHOD

### A. Data Collection and Procedure

The experimental study was conducted in a marketing and behavioural studies lab within the college campus premises in November and December 2021. In the experimental setup, the respondents were assigned to a gamified situation. Data was collected from the eye-tracking device and administered through a quantitative questionnaire. The respondents were instructed to browse through the OTT platform website, but nothing was informed about the gamification option. The participants were briefed about the various navigation choices available on the website and were allowed to browse at their leisure. In the lab, the respondents were made to sit in front of a laptop with a Tobii Eye tracker installed. After the calibration procedure, the participants explored the OTT platform website. The method of calibration involves estimating a subject's eyes' geometrical properties to create a fully-tailored and precise estimation of their gaze point location. The user is instructed to focus on specific points on the screen during calibration, also referred to as calibration dots.

In the gamified condition, the respondents were free to play games while browsing the website. The website hosted

numerous games, which included fun games, puzzles, logic games, candy games, train the brain games, and others. These games offered coins which were declared at the end. One currency was rewarded for every 10 points earned by players while playing the game. On earning coins, respondents could redeem those coins in the form of discounts or cash back. After the respondents had finished navigating through the website, they were instructed to complete the questionnaire.

### B. Participants

As the study was conducted in the institute's laboratory and respondents were approached from within the campus; thus, convenience sampling was adopted. Initially, the total number of respondents was 60. Out of these, 8 participants failed the calibration process or had an incomplete recording. Finally, 52 participants' data was used during the analysis, as given in Table II, out of which 50 percent were females, and 50 percent were males in the age group between 20 years to 35 years. The participants were presented with a memento on task completion.

TABLE II. DEMOGRAPHICS OF RESPONDENT

Items	Types	N	%
Gender	Male	26	50
	Female	26	50
Age	20-30	47	91
	30-35	5	9
Occupation	Students	52	100
Education	Post-Graduate	45	87
	Ph. D. Scholar	7	13
Time spent on the OTT platform	Less than 1 hour	7	14
	1-2 hours	20	38
	2-3 hours	25	48

Note(s): N represents the number, % represents the percentage

### C. Measures

The study variables were measured using a five-point Likert scale (Strongly Agree to Disagree Strongly) based on previous literature (see Appendix A). Individuals' interactions with the website were measured using the visualization data from eye-tracking. User engagement was measured using the user engagement scale (UES) by adopting dimensions-aesthetic appeal (AE), focused attention (FA), perceived usability (PU), and reward satisfaction (RS). It felt involvement and novelty (FN). Lastly, the study includes four control variables: age, gender, and how much time the users spend on the OTT platform daily.

### D. Attention to Watching Online Content and Gamification

Several eye-tracking statistical metrics are frequently used to translate simple eye gazing into insightful data. These metrics give a quantitative assessment and can be determined as a count, a mean, a maximum value, a minimum value, or a summation value [79]. A commonly used metric is the fixation duration and fixation count. Fixation count is used in understanding which stimuli a viewer viewed more than other

stimuli content. Another measurement, fixation duration, gives a similar meaning but with a time measurement of a stationary position of the gaze point. Reference [80] measured the visual attention of the respondent using the metrics like fixation duration, fixation counts, mouse clicks, click-through rates, and page sights. In the present study, fixation duration and fixation count are used to measure visual attention toward watching online content and gamification for analysis. For the current study based on gamification, these forms of participant-action metrics can understand the experience and awareness of the viewers.

### E. Apparatus and Material

An OTT platform service provider with a game-playing option was considered for the experiment. Tobii eye-tracking devices and software were used to quantify visual attention. The eye tracker used in the study is a hardware device clamped to the laptop screen. It is a technology that accurately records the participants' gaze behaviours. The output from the equipment is in the form of a video that shows where on the screen the participant has looked during the whole experiment. It presents visual data through gaze plots, heat maps, and clusters. The gaze points were filtered using Tobii I-VT (Attention) fixation filter [79]. The fixations duration and fixation count on the targeted options of the website were considered. Peer researchers visually reviewed a random set of coded data for quality faults.

## V. DATA ANALYSIS AND RESULTS

The data from the study were analyzed in two stages. In the first stage, visualization data for the watch (content) and play (gamification) options in the form of fixation duration and fixation count were analyzed. In the second stage, the mediation model was tested using regression analysis.

### A. First Part of the Analysis

For analysis, visualization data was collected using gaze plots and heat maps. Gaze plot data collate data for fixation duration and fixation count. The primary purpose of the gaze plot is to show the time sequence of where and when the respondent is looking at. Along with gaze plots, heatmaps are an effective tool for analyzing user behaviour on website pages, which includes user clicks, how far they scroll, and what they pay attention to or ignore. Using a heat map, researchers can identify the portion of a stimulus where participants spent the maximum time. The heatmap's color palette makes it possible to distinguish between places with longer dwell durations and shorter dwell times [81]. The heatmaps show how the total number of fixations is spread across the screen. The deeper red areas indicate which parts of the screen involved the maximum number of desires, and the green areas show the least attended function.

Fig. 2 and Fig. 3 represent aggregated heat map data images with the red color denoting maximum concentration and green indicating the least amount. At a glance, attention is more dispersed during the website navigation task, more concentrated on the play option (Fig. 2) and the watch option (Fig. 3) area on the website. Thus, stating that the visual attention towards both options was almost similar. The initial part of the quantitative finding answers the first part of the

research question- whether the viewer’s attention toward the opportunity to watch online content and the gamification option is the same. As the data under consideration (visualization data from gaze plots) has non-normal distribution, a non-parametric Mann–Whitney test was performed. The proposed hypotheses H1 and H2 were tested using the Mann-Whitney U test. This test's dependent variable is continuous or ordinal and compares variances between two independent groups.



Fig. 2. A Heat Map with Maximum Concentration on the Play Option on the OTT Platform. Source: Tobii Eye Tracker (Visualisation Data).

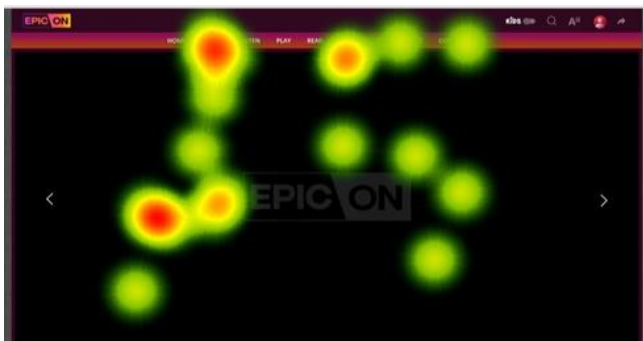


Fig. 3. A Heat Map with Maximum Concentration on the Watch Option on the OTT Platform. Source: Tobii Eye Tracker (Visualisation Data).

From Table III and Table IV, N=52, the mean rank for the option to play is more than the watch option, with U=1053.5 and p=0.052(the significance level, p<0.05) for the fixation duration and U=1056.5 and p=0.54(the significance level, p<0.05) for the fixation count, suggests that the difference between the fixation data for watch option and play option is insignificant. This states that both options' visualization data (fixation data) are not significantly different. Thus, hypotheses H1 and H2 are accepted.

TABLE III. RANKS STATISTICS FOR FIXATION DURATION (FD) AND FIXATION COUNT (FC)

Tabs	N	Fixation Duration (FD)		Fixation Count (FC)	
		Mean Rank	Sum of Ranks	Mean Rank	Sum of Ranks
Watch	52	46.76	2431.50	46.82	2434.50
Gamification	52	58.24	3028.50	58.18	3025.50
Total	104				

TABLE IV. MANN-WHITNEY TEST STATISTICS FOR FIXATION DURATION (FD) AND FIXATION COUNT (FC)

	FD	FC
Mann-Whitney U	1053.50	1056.50
Wilcoxon W	2431.50	2434.50
Z	-1.941	-1.930
Asymp. Sig. (2-tailed)	.052	.054
a. Grouping Variable: Tabs Note: p<0.05		

OTT platforms are primarily considered for viewing online content in the forms of web series, movies, and others. The above results inferred that the viewers paid equal attention to the watch option and the option to play games while browsing the website. Thus, a potential strategy for OTT platform developers to work on adopting gamification to enhance user engagement can be considered.

B. Second Part of the Analysis

In the second part of the study, the proposed hypotheses based on the mediation model were tested using regression analysis presented by [82]. The research was conducted using SPSS 22.0 software to test the mediating effect. It was examined that Reward Satisfaction (RS) acted as a mediator between attention to gamification (AG) and user engagement (UE). AG was measured using the fixation duration data from the eye-tracker. For RS and UE, data was collected using a questionnaire that had 16 items based on the User Engagement Scale (UES) [62].

1) Hypothesis testing: Previous research on gamification has incorporated regression analysis using structural equation modeling (SEM) [69]. The present study attempted to predict user engagement (dependent variable) based on attention towards gamification (independent variable) with reward satisfaction as a mediator. As recommended by [70], multiple regression analysis was used to test the model. Multiple regression analysis is an extended form of linear regression analysis. It predicts the value of a variable based on the value of two or more other variables. It determines the overall fit of the model and the contribution of each of the predictors to the total variance.

TABLE V. REGRESSION ANALYSIS

Hypothesis	R	R <sup>2</sup>	Adj.R <sup>2</sup>	F	β	P	Durbin-Watson	VI F
AG-RS	0.401	0.161	0.152	24.1	0.382	0.000	1.928	1.000
AG-UE	0.251	0.063	0.055	9.9	0.261	0.005	1.906	1.000
RS-UE	0.334	0.112	0.102	15.3	0.371	0.000	1.813	1.000

The regression analysis states that all the hypotheses have been supported as per Table V. Durbin-Watson statistics is between 1.5 and 2.5; hence there is no autocorrelation [83]. By analyzing the path connecting AG to RS (H4), it can be stated that AG substantially impacts RS for the viewers. It explains

nearly 15.2% of the total RS variance. The  $\beta$  coefficient of this path is 0.382 and was found to be statistically significant at  $p < 0.000$ .

The mediating regression analysis output is presented in Table VI. The statistics propose that AG under the mediating effect of RS (H6) is an essential predictor of UE. Sobel statistics suggested by [84], with a  $p$ -value  $< 0.05$ , was used to signify the mediating regression. The study revealed that the direct effect ( $c' = 0.15$ ) of AG on UE is insignificant, whereas the indirect effect ( $a \times b = 0.124$ ,  $p < 0.05$ ) is significant, indicating that RS plays a mediating role between AG and UE. Complete mediation occurs in a condition where the independent variable has no effect when the mediator is controlled [82]. AG accounted for 5% variability in UE. However, when RS was introduced as a mediating variable between AG and UE, the variability increased to 12%, which is more than double compared to the AG- UE model. Thus, the model improved with the mediating effect of RS.

Our study highlights that gamification adoption improves user engagement with rewards as one of the key contributors. Therefore, the developers need to pay attention to innovating and improving the adoption of gamification across OTT platforms. This seems logical because when the user is provided with an option to play and earn, they feel motivated and engage on the platform. Although the application of gamification on the OTT service providers' platform is in its primary stage, our study highlights a positive future. Considering that the satisfaction of the reward is a mediator, it will be a practical step to improvise the quality of the reward to improve user engagement. This improvement seems reasonable as the users will have options to earn good quality rewards while playing short and easy games on the OTT platform [11], [85]. It will enhance their interest in playing and watching online content.

TABLE VI. MEDIATING REGRESSION ANALYSIS

IV	MV	DV	Effects of IV on MV (a)	Effects of MV on DV (b)	Direct effect (c')	Indirect effect (a × b)	Total effects c'+(a × b)	Mediation	Sobel p-value
AG	RS	UE	0.382 (0.087)*	0.371 (0.092)*	0.15 (0.09)	0.124	0.261 (0.091)	Complete mediation	0.002

Notes: IV, independent variable; MV, mediating variable; DV, dependent variable; AG, Attention to gamification; RS, reward satisfaction; UE, User Engagement.  
\* $p < 0.05$

## VI. DISCUSSION AND IMPLICATIONS OF RESEARCH

Many companies are driving their focus toward implementation gamification in their business processes to improve employee-customer engagement. Businesses employ game features to increase repeat purchasing behaviour. This study aims to understand if the adoption of gamification with reward enhances user engagement. While previous studies [69], [86] have attempted to link gamification to customer engagement, the current study mediation process was used to understand how rewards motivate viewers on the OTT platform. This part of the paper suggests theoretical and practical applications of the study conducted, thereby adding to the body of knowledge by advancing the theoretical and practical discussion related to gamification adoption and behavioural studies.

This study makes several significant theoretical contributions. Firstly, the study is one of its kind to demonstrate the relation between gamification, reward, and user engagement. An experiment was designed to test how gamification and rewards influenced viewer engagement. Earlier studies have used the gamification concept across platforms like tourism websites, mobile apps, crowdfunding platforms, and others. In contrast, the current study considered an OTT platform to test the impact of gamification on viewer engagement. Since its inception, users have been using OTT platforms to watch online content in web series, movies, and others. This study brings a new perspective toward understanding the role of gamification on the OTT platform, which is in its booming stage. Previous studies concerning OTT platforms have focused on understanding the users' adoption behaviour. With the adoption of gamification on the

OTT platform, OTT service providers are reshaping their basic structures of providing online content. Thus, this study provides insight into the practicality of the OTT platform's new approach, like gamification. A survey by [87] stated that reward was a mediator in the relationship between gamification and motivation (hedonic and utilitarian). Similarly, the current study successfully analyses the mediating role of rewards in the relationship between gamification and user engagement. This study is a first-hand approach to conducting eye-tracking analysis on the OTT platform, thereby adding to the literature on the methods used across OTT platforms.

In addition to the theoretical contributions, several managerial implications derive from the study's empirical results. Integrating gamification components into marketing, social media, community, and other digital brand experiences is a potent user engagement tactic for increasing engagement. Though not a one-size-fits-all solution, gamification can assist in increasing brand and content awareness in the OTT video market in various ways when used carefully and tailored to the content [88]. Earlier research demonstrated that gamification positively impacts buying behaviour and engagement [54], [61]. The results from the present study support the inclusion of gamification for increasing engagement. The findings state several practical suggestions to help OTT platform developers and marketers make better decisions regarding the implementation of gamification. With gamification getting a similar amount of visual attention compared to watching content options, marketers and developers must work hard towards the appropriate implementation of gamification to increase user engagement on OTT platforms.

Game components like rewards (received upon playing games) make it desirable for consumers to interact with



gamification options and influence user engagement. Game designers can think about including challenges and real-time feedback in their games so that users can keep track of their improvement and outcomes. Players can receive points for their efforts and, based on the points earned, progress to advanced levels with more challenging activities, giving them the impression that their abilities are growing and that spending time on such platforms is valuable. Gamification on the OTT platform is in the nascent stage, so game developers should pay attention to the game's aesthetic appeal as it is the first thing the user views. Games that are visually attractive with high-quality graphics have a high chance of getting spotted.

The present study used an eye-tracking device to map the OTT viewers' visual attention and platform activities. Eye-tracking (Eye movement analysis) is a valuable source for investigating the visual attention process (which is directed by the sub-conscious behaviour) while involved in an activity. It displays how visual attention is allocated to the objects (e.g., words and graphic portions), how long, and in what direction [68]. The eye-tracking approach has been used widely in library science and information search [89]. This study has used eye-tracking software to analyze user engagement through visual attention data towards gamification on the OTT platform and is one of a kind.

#### VII. LIMITATIONS AND FUTURE RESEARCH

The study provides possibilities for further research. The study adopted a unique approach to understanding the impact of gamification on OTT with a preference for user engagement. Although this study makes several contributions, it also has several limitations. The information was gathered all at once; longitudinal data could be used in future studies to determine the long-term impact of gamification. Second, the data was collected via a single OTT platform; thus, future studies might be conducted on different OTT. Thirdly, future research could carry out a comparative study between the two or more OTT platforms with or without gamification.

In the current study, the sample size was limited to only 52. The respondent was college campus students with an average age between 20 to 30 years; future research could use an improved sample size and age range above 35 years. With respondents from the Millennials category, the interest, involvement, and engagement could give mixed results. The present study was purely quantitative; in the future, mixed-method or qualitative research might generate different insights. Also, the future progress in this study could be to understand the continued usage intention [90] driven by user engagement towards the OTT platform. With the OTT content becoming part of everyday life, it has become a topic of conversation among viewers. Finding, watching, and discussing content has become a social experience amongst online communities. OTT providers must comprehend how potential viewers find specific and exciting content and apply this knowledge to their marketing strategies, generating a requirement for future research to explore social interaction and engagement on the OTT platform.

#### VIII. CONCLUSION

The result of this study showcases how gamification affects users while watching online content on the OTT platform. With the intrinsically motivating aspect of the gamified services [91], usage of gamification on the OTT platforms could increase due to its appealing approach towards the users. The present study does not engage with the general inclination to use the app; it only shows how the implementation of gamification can improve the viewer's engagement on the OTT platform and can be used as an efficient marketing tool for promoting the OTT apps.

#### REFERENCES

- [1] "Over The Top (OTT) Services Market," 2021. [Online]. Available: <https://www.fortunebusinessinsights.com/industry-reports/over-the-top-services-market-100506>.
- [2] S.-H. Yoon, H.-W. Kim, and A. Kankanhalli, "What makes people watch online TV clips? An empirical investigation of survey data and viewing logs," *Int. J. Inf. Manage.*, vol. 59, p. 102329, Aug. 2021, doi: 10.1016/j.ijinfomgt.2021.102329.
- [3] G. Gupta and K. Singharia, "Consumption of OTT Media Streaming in COVID-19 Lockdown: Insights from PLS Analysis," *Vis. J. Bus. Perspect.*, vol. 25, no. 1, pp. 36–46, Mar. 2021, doi: 10.1177/0972262921989118.
- [4] T. YS, "ZEE5, India's largest OTT platform reimagines Hyper-personalisation for consumers," ZEE5 View Brand Publisher, 2021.
- [5] A. Bozkurt and G. Durak, "A Systematic Review of Gamification Research," *Int. J. Game-Based Learn.*, vol. 8, no. 3, pp. 15–33, Jul. 2018, doi: 10.4018/IJGBL.2018070102.
- [6] A. Domínguez, J. Saenz-de-Navarrete, L. De-Marcos, L. Fernández-Sanz, C. Pagés, and J.-J. Martínez-Herráiz, "Gamifying learning experiences: Practical implications and outcomes," *Comput. Educ.*, vol. 63, pp. 380–392, Apr. 2013, doi: 10.1016/j.compedu.2012.12.020.
- [7] K. Huotari and J. Hamari, "A definition for gamification: anchoring gamification in the service marketing literature," *Electron. Mark.*, vol. 27, no. 1, pp. 21–31, 2017, doi: 10.1007/s12525-015-0212-z.
- [8] J. Koivisto and J. Hamari, "The rise of motivational information systems: A review of gamification research," *Int. J. Inf. Manage.*, vol. 45, pp. 191–210, Apr. 2019, doi: 10.1016/j.ijinfomgt.2018.10.013.
- [9] Y. Yang, Y. Asaad, and Y. Dwivedi, "Examining the impact of gamification on intention of engagement and brand attitude in the marketing context," *Comput. Human Behav.*, vol. 73, pp. 459–469, 2017, doi: 10.1016/j.chb.2017.03.066.
- [10] C. F. Hofacker, K. de Ruyter, N. H. Lurie, P. Manchanda, and J. Donaldson, "Gamification and Mobile Marketing Effectiveness," *J. Interact. Mark.*, vol. 34, pp. 25–36, May 2016, doi: 10.1016/j.intmar.2016.03.001.
- [11] J. Högberg, P. Shams, and E. Wästlund, "Gamified in-store mobile marketing: The mixed effect of gamified point-of-purchase advertising," *J. Retail. Consum. Serv.*, vol. 50, no. July 2018, pp. 298–304, 2019, doi: 10.1016/j.jretconser.2018.07.004.
- [12] D. M.-H. Wen, D. J.-W. Chang, Y.-T. Lin, C.-W. Liang, and S.-Y. Yang, "Gamification Design for Increasing Customer Purchase Intention in a Mobile Marketing Campaign App," 2014, pp. 440–448.
- [13] J. Liu, X. Zhang, F. Meng, and K. Lai, "Deploying gamification to engage physicians in an online health community: An operational paradox," *Int. J. Prod. Econ.*, vol. 228, p. 107847, Oct. 2020, doi: 10.1016/j.ijpe.2020.107847.
- [14] S. Nagaraj, S. Singh, and V. R. Yasa, "Factors affecting consumers' willingness to subscribe to over-the-top (OTT) video streaming services in India," *Technol. Soc.*, vol. 65, p. 101534, May 2021, doi: 10.1016/j.techsoc.2021.101534.
- [15] M. L. Wayne and D. Castro, "SVOD Global Expansion in Cross-National Comparative Perspective: Netflix in Israel and Spain," *Telev. New Media*, vol. 22, no. 8, pp. 896–913, Dec. 2021, doi: 10.1177/1527476420926496.

- [16] S. Deterding, "Situational motivational affordances of game elements: A conceptual model," CHI 2011 Work. "Gamification," no. July, 2011, [Online]. Available: <http://gamification-research.org/chi2011/papers>.
- [17] G. Barata, S. Gama, J. Jorge, and D. Gonçalves, "Gamification for smarter learning: tales from the trenches," Smart Learn. Environ., vol. 2, no. 1, p. 10, Dec. 2015, doi: 10.1186/s40561-015-0017-8.
- [18] G. Lucassen and S. Jansen, "Gamification in Consumer Marketing - Future or Fallacy?," Procedia - Soc. Behav. Sci., vol. 148, pp. 194–202, Aug. 2014, doi: 10.1016/j.sbspro.2014.07.034.
- [19] T. Leclercq, W. Hammadi, and I. Poncin, "The Boundaries of Gamification for Engaging Customers: Effects of Losing a Contest in Online Co-creation Communities," J. Interact. Mark., vol. 44, pp. 82–101, Nov. 2018, doi: 10.1016/j.intmar.2018.04.004.
- [20] R. Pace and A. Dipace, "Game-Based Learning and Lifelong Learning for Tourist Operators," in Cultural Tourism in Digital Era, Springer International Publishing, 2015, pp. 185–199.
- [21] D. Pacauskas, R. Rajala, M. Westerlund, and M. Mäntymäki, "Harnessing user innovation for social media marketing: Case study of a crowdsourced hamburger," Int. J. Inf. Manage., vol. 43, pp. 319–327, Dec. 2018, doi: 10.1016/j.ijinfomgt.2018.08.012.
- [22] J. Müller-Stewens, T. Schlager, G. Häubl, and A. Herrmann, "Gamified Information Presentation and Consumer Adoption of Product Innovations," J. Mark., vol. 81, no. 2, pp. 8–24, Mar. 2017, doi: 10.1509/jm.15.0396.
- [23] F. Espigares-Jurado, F. Muñoz-Leiva, M. B. Correia, C. M. R. Sousa, C. M. Q. Ramos, and L. Fáisca, "Visual attention to the main image of a hotel website based on its position, type of navigation and belonging to Millennial generation: An eye tracking study," J. Retail. Consum. Serv., vol. 52, p. 101906, Jan. 2020, doi: 10.1016/j.jretconser.2019.101906.
- [24] T. Harwood and T. Garry, "An investigation into gamification as a customer engagement experience environment," J. Serv. Mark., vol. 29, no. 6/7, pp. 533–546, Sep. 2015, doi: 10.1108/JSM-01-2015-0045.
- [25] H. L. O'Brien and E. G. Toms, "What is user engagement? A conceptual framework for defining user engagement with technology," J. Am. Soc. Inf. Sci. Technol., vol. 59, no. 6, pp. 938–955, Apr. 2008, doi: 10.1002/asi.20801.
- [26] R. S. Alsawaier, "The effect of gamification on motivation and engagement," Int. J. Inf. Learn. Technol., vol. 35, no. 1, pp. 56–79, Jan. 2018, doi: 10.1108/IJILT-02-2017-0009.
- [27] J. Bowden, "Customer Engagement: A Framework for Assessing Customer-Brand Relationships: The Case of the Restaurant Industry," J. Hosp. Mark. Manag., vol. 18, no. 6, pp. 574–596, Jul. 2009, doi: 10.1080/19368620903024983.
- [28] J. L.-H. Bowden, "The Process of Customer Engagement: A Conceptual Framework," J. Mark. Theory Pract., vol. 17, no. 1, pp. 63–74, Jan. 2009, doi: 10.2753/MTP1069-6679170105.
- [29] L. Hollebeek, "Exploring customer brand engagement: definition and themes," J. Strateg. Mark., vol. 19, no. 7, pp. 555–573, Dec. 2011, doi: 10.1080/0965254X.2011.599493.
- [30] R. Eppmann, M. Bekk, K. Klein, and F. Völkner, "Understanding the (negative and positive) effects of gamification for companies," Mark. Sci. Inst. Work. Pap. Ser., vol. 20–14, pp. 200–212, 2020, [Online]. Available: [file:///C:/Users/marvi/Downloads/study\\_id27102\\_cash-and-carry-grosshandel-in-deutschland-statista-dossier.pdf%0Ahttps://de.statista.com/statistik/studie/id/25139/dokument/lebensmittelhandel-in-oesterreich-statista-dossier/%0Ahttp://www.lebensmittelzeitung](file:///C:/Users/marvi/Downloads/study_id27102_cash-and-carry-grosshandel-in-deutschland-statista-dossier.pdf%0Ahttps://de.statista.com/statistik/studie/id/25139/dokument/lebensmittelhandel-in-oesterreich-statista-dossier/%0Ahttp://www.lebensmittelzeitung).
- [31] C. S. L. Tan, "Gamifying OTT: a study on consumer attitudes toward game elements and OTT media service provider brands in gamification," Young Consum., vol. 22, no. 3, pp. 328–347, Jul. 2021, doi: 10.1108/YC-11-2020-1245.
- [32] B. Hutchins, B. Li, and D. Rowe, "Over-the-top sport: live streaming services, changing coverage rights markets and the growth of media sport portals," Media, Cult. Soc., vol. 41, no. 7, pp. 975–994, Oct. 2019, doi: 10.1177/0163443719857623.
- [33] K. T. Kwak, C. J. Oh, and S. W. Lee, "Who uses paid over-the-top services and why? Cross-national comparisons of consumer demographics and values," Telecomm. Policy, vol. 45, no. 7, p. 102168, Aug. 2021, doi: 10.1016/j.telpol.2021.102168.
- [34] R. L. Oliver, "A Cognitive Model of the Antecedents and Consequences of Satisfaction Decisions," J. Mark. Res., vol. 17, no. 4, pp. 460–469, Nov. 1980, doi: 10.1177/002224378001700405.
- [35] F. D. Davis, "Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology," MIS Q., vol. 13, no. 3, p. 319, Sep. 1989, doi: 10.2307/249008.
- [36] Van Der Heijden, "User Acceptance of Hedonic Information Systems," MIS Q., vol. 28, no. 4, p. 695, 2004, doi: 10.2307/25148660.
- [37] H. Choi and S. Kim, "국내외 OTT 서비스의 사용자 경험 연구," vol. 18, no. 4, pp. 425–431, 2020.
- [38] G. Sahu, L. Gaur, and G. Singh, "Applying niche and gratification theory approach to examine the users' indulgence towards over-the-top platforms and conventional TV," Telemat. Informatics, vol. 65, p. 101713, Dec. 2021, doi: 10.1016/j.tele.2021.101713.
- [39] S. Kim, H. Baek, and D. H. Kim, "OTT and live streaming services: Past, present, and future," Telecomm. Policy, vol. 45, no. 9, p. 102244, Oct. 2021, doi: 10.1016/j.telpol.2021.102244.
- [40] H. Syrjälä, H. Kauppinen-Räsänen, H. T. Luomala, T. N. Joëlsson, K. Könnölä, and T. Mäkilä, "Gamified package: Consumer insights into multidimensional brand engagement," J. Bus. Res., vol. 119, no. December 2019, pp. 423–434, 2020, doi: 10.1016/j.jbusres.2019.11.089.
- [41] A. Pansari and V. Kumar, "Customer engagement: the construct, antecedents, and consequences," J. Acad. Mark. Sci., vol. 45, no. 3, pp. 294–311, May 2017, doi: 10.1007/s11747-016-0485-6.
- [42] I. Bouchrika, N. Harrati, V. Wanick, and G. Wills, "Exploring the impact of gamification on student engagement and involvement with e-learning systems," Interact. Learn. Environ., vol. 29, no. 8, pp. 1244–1257, Nov. 2021, doi: 10.1080/10494820.2019.1623267.
- [43] K. Robson, K. Plangger, J. H. Kietzmann, I. McCarthy, and L. Pitt, "Game on: Engaging customers and employees through gamification," Bus. Horiz., vol. 59, no. 1, pp. 29–36, Jan. 2016, doi: 10.1016/j.bushor.2015.08.002.
- [44] W. Hammadi, T. Leclercq, and A. C. R. Van Riel, "The use of gamification mechanics to increase employee and user engagement in participative healthcare services," J. Serv. Manag., vol. 28, no. 4, pp. 640–661, Aug. 2017, doi: 10.1108/JOSM-04-2016-0116.
- [45] Y. Yang, Y. Asaad, and Y. Dwivedi, "Examining the impact of gamification on intention of engagement and brand attitude in the marketing context," Comput. Human Behav., vol. 73, pp. 459–469, Aug. 2017, doi: 10.1016/j.chb.2017.03.066.
- [46] F. Xu, D. Buhalis, and J. Weber, "Serious games and the gamification of tourism," Tour. Manag., vol. 60, pp. 244–256, Jun. 2017, doi: 10.1016/j.tourman.2016.11.020.
- [47] S. Jang, P. J. Kitchen, and J. Kim, "The effects of gamified customer benefits and characteristics on behavioral engagement and purchase: Evidence from mobile exercise application uses," J. Bus. Res., vol. 92, pp. 250–259, Nov. 2018, doi: 10.1016/j.jbusres.2018.07.056.
- [48] A. B. Eisingerich, A. Marchand, M. P. Fritze, and L. Dong, "Hook vs. hope: How to enhance customer engagement through gamification," Int. J. Res. Mark., vol. 36, no. 2, pp. 200–215, Jun. 2019, doi: 10.1016/j.ijresmar.2019.02.003.
- [49] A. Berger, T. Schlager, D. E. Sprott, and A. Herrmann, "Gamified interactions: whether, when, and how games facilitate self-brand connections," J. Acad. Mark. Sci., vol. 46, no. 4, pp. 652–673, Jul. 2018, doi: 10.1007/s11747-017-0530-0.
- [50] C.-L. Hsu and M.-C. Chen, "How does gamification improve user experience? An empirical investigation on the antecedences and consequences of user experience and its mediating role," Technol. Forecast. Soc. Change, vol. 132, pp. 118–129, Jul. 2018, doi: 10.1016/j.techfore.2018.01.023.
- [51] C.-L. Hsu and M.-C. Chen, "How gamification marketing activities motivate desirable consumer behaviors: Focusing on the role of brand love," Comput. Human Behav., vol. 88, pp. 121–133, Nov. 2018, doi: 10.1016/j.chb.2018.06.037.
- [52] N. Xi and J. Hamari, "Does gamification affect brand engagement and equity? A study in online brand communities," J. Bus. Res., vol. 109, pp. 449–460, Mar. 2020, doi: 10.1016/j.jbusres.2019.11.058.

- [53] M.-S. Kuo and T.-Y. Chuang, "How gamification motivates visits and engagement for online academic dissemination – An empirical study," *Comput. Human Behav.*, vol. 55, pp. 16–27, Feb. 2016, doi: 10.1016/j.chb.2015.08.025.
- [54] X. Wang, D. H.-L. Goh, E.-P. Lim, A. W. L. Vu, and A. Y. K. Chua, "Examining the Effectiveness of Gamification in Human Computation," *Int. J. Human-Computer Interact.*, vol. 33, no. 10, pp. 813–821, Oct. 2017, doi: 10.1080/10447318.2017.1287458.
- [55] A. Suh, C. Wagner, and L. Liu, "Enhancing User Engagement through Gamification," *J. Comput. Inf. Syst.*, vol. 58, no. 3, pp. 204–213, Jul. 2018, doi: 10.1080/08874417.2016.1229143.
- [56] N. P. Cechetti, E. A. Bellei, D. Biduski, J. P. M. Rodriguez, M. K. Roman, and A. C. B. De Marchi, "Developing and implementing a gamification method to improve user engagement: A case study with an m-Health application for hypertension monitoring," *Telemat. Informatics*, vol. 41, pp. 126–138, Aug. 2019, doi: 10.1016/j.tele.2019.04.007.
- [57] M. Featherstone and J. Habgood, "UniCraft: Exploring the impact of asynchronous multiplayer game elements in gamification," *Int. J. Hum. Comput. Stud.*, vol. 127, pp. 150–168, Jul. 2019, doi: 10.1016/j.ijhcs.2018.05.006.
- [58] W. Feng, R. Tu, and P. Hsieh, "Can gamification increase consumers' engagement in fitness apps? The moderating role of commensurability of the game elements," *J. Retail. Consum. Serv.*, vol. 57, p. 102229, Nov. 2020, doi: 10.1016/j.jretconser.2020.102229.
- [59] S. Kamboj, S. Rana, and V. A. Drave, "Factors Driving Consumer Engagement and Intentions with Gamification of Mobile Apps," *J. Electron. Commer. Organ.*, vol. 18, no. 2, pp. 17–35, Apr. 2020, doi: 10.4018/JECO.2020040102.
- [60] N. Xi and J. Hamari, "Does gamification satisfy needs? A study on the relationship between gamification features and intrinsic need satisfaction," *Int. J. Inf. Manage.*, vol. 46, no. November 2018, pp. 210–221, 2019, doi: 10.1016/j.ijinfomgt.2018.12.002.
- [61] M. Silic, G. Marzi, A. Caputo, and P. M. Bal, "The effects of a gamified human resource management system on job satisfaction and engagement," *Hum. Resour. Manag. J.*, vol. 30, no. 2, pp. 260–277, Apr. 2020, doi: 10.1111/1748-8583.12272.
- [62] H. L. O'Brien and E. G. Toms, "The development and evaluation of a survey to measure user engagement," *J. Am. Soc. Inf. Sci. Technol.*, vol. 61, no. 1, pp. 50–69, Jan. 2010, doi: 10.1002/asi.21229.
- [63] H. L. O'Brien, P. Cairns, and M. Hall, "A practical approach to measuring user engagement with the refined user engagement scale (UES) and new UES short form," *Int. J. Hum. Comput. Stud.*, vol. 112, no. July 2017, pp. 28–39, 2018, doi: 10.1016/j.ijhcs.2018.01.004.
- [64] T. Blascheck, K. Kurzhals, M. Raschke, M. Burch, D. Weiskopf, and T. Ertl, "Visualization of Eye Tracking Data: A Taxonomy and Survey," *Comput. Graph. Forum*, vol. 36, no. 8, pp. 260–284, Dec. 2017, doi: 10.1111/cgf.13079.
- [65] A. Schall and J. Romano Bergstrom, "The Future of Eye Tracking and User Experience," in *Eye Tracking in User Experience Design*, Elsevier, 2014, pp. 351–360.
- [66] B. W. Wojdyski and H. Bang, "Distraction effects of contextual advertising on online news processing: an eye-tracking study," *Behav. Inf. Technol.*, vol. 35, no. 8, pp. 654–664, Aug. 2016, doi: 10.1080/0144929X.2016.1177115.
- [67] P. C. Heather O'Brien, *Why engagement matters: Cross-disciplinary perspectives of user engagement in digital media*. Springer International Publishing, 2016.
- [68] T.-Y. Kim and D.-H. Shin, "User Experience(UX) of Facebook: Focusing on Users' Eye Movement Pattern and Advertising Contents," *J. Korea Contents Assoc.*, vol. 14, no. 7, pp. 45–57, Jul. 2014, doi: 10.5392/JKCA.2014.14.07.045.
- [69] A. García-Jurado, M. Torres-Jiménez, A. L. Leal-Rodríguez, and P. Castro-González, "Does gamification engage users in online shopping?," *Electron. Commer. Res. Appl.*, vol. 48, p. 101076, Jul. 2021, doi: 10.1016/j.elerap.2021.101076.
- [70] J. Hamari, J. Koivisto, and H. Sarsa, "Does Gamification Work? -- A Literature Review of Empirical Studies on Gamification," in *2014 47th Hawaii International Conference on System Sciences*, Jan. 2014, pp. 3025–3034, doi: 10.1109/HICSS.2014.377.
- [71] J. Högberg, J. Hamari, and E. Wästlund, "Gameful Experience Questionnaire (GAMEFULQUEST): an instrument for measuring the perceived gamefulness of system use," *User Model. User-adapt. Interact.*, vol. 29, no. 3, pp. 619–660, Jul. 2019, doi: 10.1007/s11257-019-09223-w.
- [72] G. Loewenstein and S. Issacharoff, "Source dependence in the valuation of objects," *J. Behav. Decis. Mak.*, vol. 7, no. 3, pp. 157–168, Sep. 1994, doi: 10.1002/bdm.3960070302.
- [73] L. F. Rodrigues, A. Oliveira, and H. Rodrigues, "Main gamification concepts: A systematic mapping study," *Heliyon*, vol. 5, no. 7, p. e01993, Jul. 2019, doi: 10.1016/j.heliyon.2019.e01993.
- [74] J. Hwang and L. Choi, "Having fun while receiving rewards?: Exploration of gamification in loyalty programs for consumer loyalty," *J. Bus. Res.*, vol. 106, pp. 365–376, Jan. 2020, doi: 10.1016/j.jbusres.2019.01.031.
- [75] E. Marcucci, V. Gatta, and M. Le Pira, "Gamification design to foster stakeholder engagement and behavior change: An application to urban freight transport," *Transp. Res. Part A Policy Pract.*, vol. 118, pp. 119–132, Dec. 2018, doi: 10.1016/j.tra.2018.08.028.
- [76] Y. (Sandy) Shen, H. C. Choi, M. Joppe, and S. Yi, "What motivates visitors to participate in a gamified trip? A player typology using Q methodology," *Tour. Manag.*, vol. 78, p. 104074, Jun. 2020, doi: 10.1016/j.tourman.2019.104074.
- [77] C. Jennett et al., "Measuring and defining the experience of immersion in games," *Int. J. Hum. Comput. Stud.*, vol. 66, no. 9, pp. 641–661, Sep. 2008, doi: 10.1016/j.ijhcs.2008.04.004.
- [78] R. M. Ryan and E. L. Deci, "Intrinsic and Extrinsic Motivations: Classic Definitions and New Directions," *Contemp. Educ. Psychol.*, vol. 25, no. 1, pp. 54–67, Jan. 2000, doi: 10.1006/ceps.1999.1020.
- [79] A. Olsen, "The Tobii I-VT Fixation Filter," 2012. [Online]. Available: <https://www.tobii.com/siteassets/tobii-pro/learn-and-support/analyze/how-do-we-classify-eye-movements/tobii-pro-i-vt-fixation-filter.pdf>.
- [80] R. M. Martey et al., "Measuring Game Engagement," *Simul. Gaming*, vol. 45, no. 4–5, pp. 528–547, Aug. 2014, doi: 10.1177/1046878114553575.
- [81] A. T. Duchowski, *Eye Tracking Methodology*. Cham: Springer International Publishing, 2017.
- [82] R. M. Baron and D. A. Kenny, "The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations," *J. Pers. Soc. Psychol.*, vol. 51, no. 6, pp. 1173–1182, 1986, doi: 10.1037/0022-3514.51.6.1173.
- [83] J. F. Hair Jr., L. M. Matthews, R. L. Matthews, and M. Sarstedt, "PLS-SEM or CB-SEM: updated guidelines on which method to use," *Int. J. Multivar. Data Anal.*, vol. 1, no. 2, p. 107, 2017, doi: 10.1504/ijmda.2017.10008574.
- [84] M. E. Sobel, "Asymptotic Confidence Intervals for Indirect Effects in Structural Equation Models," *Sociol. Methodol.*, vol. 13, p. 290, 1982, doi: 10.2307/270723.
- [85] P. Bitrián, I. Buil, and S. Catalán, "Enhancing user engagement: The role of gamification in mobile apps," *J. Bus. Res.*, vol. 132, no. April, pp. 170–185, 2021, doi: 10.1016/j.jbusres.2021.04.028.
- [86] S. Shi, W. K. S. Leung, and F. Munelli, "Gamification in OTA platforms: A mixed-methods research involving online shopping carnival," *Tour. Manag.*, vol. 88, p. 104426, Feb. 2022, doi: 10.1016/j.tourman.2021.104426.
- [87] A. Behl and V. Pereira, "What's behind a scratch card? Designing a mobile application using gamification to study customer loyalty: An experimental approach," *Australas. J. Inf. Syst.*, vol. 25, Nov. 2021, doi: 10.3127/ajis.v25i0.3203.
- [88] M. Khurana, "MOBILE ECOSYSTEM REPORT 2020," 2020. [Online]. Available: [https://d2ksis2z2ke2jq.cloudfront.net/uploads/2020/02/mmer\\_2020\\_-\\_report.pdf](https://d2ksis2z2ke2jq.cloudfront.net/uploads/2020/02/mmer_2020_-_report.pdf).
- [89] J. Walhout, P. Oomen, H. Jarodzka, and S. Brand-Gruwel, "Effects of task complexity on online search behavior of adolescents," *J. Assoc. Inf. Sci. Technol.*, vol. 68, no. 6, pp. 1449–1461, Jun. 2017, doi: 10.1002/asi.23782.
- [90] R. Pereira and C. Tam, "Impact of enjoyment on the usage continuance intention of video-on-demand services," *Inf. Manag.*, vol. 58, no. 7, p. 103501, Nov. 2021, doi: 10.1016/j.im.2021.103501.

[91] R. S. Alsawaier, "Research trends in the study of gamification," Int. J. Inf. Learn. Technol., vol. 36, no. 5, pp. 373–380, 2019, doi: 10.1108/IJILT-12-2017-0119.

[92] H. L. O'Brien, P. Cairns, and M. Hall, "A practical approach to measuring user engagement with the refined user engagement scale (UES) and new UES short form," Int. J. Hum. Comput. Stud., vol. 112, pp. 28–39, Apr. 2018, doi: 10.1016/j.ijhcs.2018.01.004.

APPENDIX- A

Construct and source		Items
User Engagement [92]	AE.1	The games on the OTT platform were attractive
	AE.2	The components of the game on the OTT platform were aesthetically appealing.
	AE.3	I liked the graphics and images of the game on the OTT platform.
	FA.1	While playing the game on the OTT platform, I lost myself.
	FA.2	I was so involved in playing the game on OTT that I lost track of time.
	FA.3	I blocked out things around me when I was playing the game on the OTT platform.
	PU.1	I felt frustrated while using the OTT platform for playing games. (R)
	PU.2	I found the game on the OTT platform confusing to use. (R)
	PU.3	I felt annoyed while playing the game on the OTT platform. (R)
	NO1	I continued to play games on the OTT platform out of curiosity.
	FI2	I felt involved in the gaming task.
	FI3	This gaming experience on the OTT platform was fun.
	RW.1	Using the OTT platform to play games was worthwhile.
	RW.3	The experience of playing a game on the OTT platform did not work out the way I had planned.
	RW.4	My experience while playing games on the OTT platform was rewarding.
RW.5	I recommend playing a game on the OTT platform to my family and friends.	

# Optimally Allocating Ambulances in Delhi using Mutation based Shuffled Frog Leaping Algorithm

Zaheeruddin<sup>1</sup>, Hina Gupta<sup>2\*</sup>

Department of Electrical Engineering  
Jamia Millia Islamia University, New Delhi, India

**Abstract**—This paper presents a reliable and competent evolutionary-based approach for improving the response time of Emergency Medical Service (EMS) by efficiently allocating ambulances at the base stations. As the prime objective of EMS is to save people's lives by providing them with timely assistance, thus increasing the chances of a person's survivability, this paper has undertaken the problem of ambulance allocation. The work has been implemented using the proposed mutation-based Shuffled Frog Leaping Algorithm (mSFLA) to provide an optimal allocation plan. The authors have altered the basic SFLA using the concept of mutation to improve the quality of the solution obtained and avoid being trapped in local optima. Considering a set of assumptions, the new algorithm has been applied for allocating 50 ambulances among 11 base stations in Southern Delhi. The working environment of EMS, which includes stochastic requests, travel time, and dynamic traffic conditions, has been considered to attain accurate results. The work has been implemented in the MATLAB simulation environment to find an optimized allocation plan with a minimum average response time. The authors have reduced the average response time by 12.23% with the proposed algorithm. The paper also compares mSFLA, Genetic Algorithm (GA), and Particle Swarm Optimization (PSO) for the stated problem. The algorithms are compared in terms of objective value (average response time), convergence rate, and constancy repeatability to conclude that mSFLA performs better than the other two algorithms.

**Keywords**—Ambulance allocation; ambulance service; emergency medical service; shuffled frog leaping algorithm; mutation based shuffled frog leaping algorithm

## I. INTRODUCTION

Emergency Medical Service (EMS) control centers are vital in modern health systems and act as a pre-hospital component. EMSs provide out-of-hospital medical care and transport activities for the victims of accidents or illnesses. It plays a significant role in the public health system, and its ability to respond to emergency calls can significantly impact a patient's health and recovery [1]. Therefore, EMS control centers need to strategize and work towards handling a significant concern of allocating an appropriate count of ambulances to the base stations in an area [2]. Having a suitable count of ambulances available at the base stations will help the EMS provide a timely response to the persons in need. This motive has attracted many researchers to suggest solutions that prove viable in the working environment of EMS. As per the working procedure, an ambulance is selected and dispatched to the demand site when the EMS control center receives a request call. The rules set by the EMS authority help select and

dispatch the ambulance from one of the base stations (with ambulance availability) to serve the request. The rule may select the nearest ambulance to the requested location or the ambulance that will take less time to reach the location [3]–[5]. When the ambulance reaches the requested location, it may provide first aid to the patient or resuscitation. It then takes the patient to the hospital as per the requirement.

Research carried out to date has emphasized many issues related to planning, working, and management activities related to EMS using static models, dynamic models, hypercube models, covering models, etc. To attain optimum service performance, EMS facilities must be positioned strategically in a specific locality. Decisions here are taken from two aspects: selecting appropriate sites at which ambulances should be stationed and the number of ambulances stationed at each site [6]. However, considering densely populated cities and countries, deciding on allotting and constructing new places for base stations is challenging for the government. Therefore, the authors have undertaken the problem of optimizing the performance of EMSs by finding an optimal allocation solution for distributing ambulances among the existing base stations using the details regarding (1) the number of ambulances in the fleet, (2) the location of base stations, (3) frequency of request calls, and (4) tentative demand sites.

The remaining paper is organized as follows: Section II covers the literature review; Section III focuses on the problem background; followed by problem formulation in Section IV. Simulation modeling has been explained in Section V. Section VI covers the details related to simulation, results obtained, and discussions related to the same. Finally, the paper is concluded in Section VII.

## II. LITERATURE REVIEW

Out of all the literature available on this domain, the authors have cited works primarily focusing on optimizing ambulance allocation. Ambulance allocation is the distribution of ambulances to the base stations based on specific criteria [7]. The base station is an area where the ambulances are in standby mode. They are dispatched from the base station when they have to serve any request. However, deciding the locations for positioning the base station is out of the scope of this work. Our work primarily focuses on finding the allocation plan for ambulances at the existing base stations. Many researchers have extensively explored the EMS field to improve the service level provided to society. The research in the field of EMS needs the answer to the following two queries: (1) the optimization criterion that can be used as the

\*Corresponding Author.

best proxy for health outcomes, and (2) the model that is apt to be used for designing EMS systems handling a mixed territory comprising urban and rural areas. The work of many authors has covered the answer to the first query. In work put forward by [8]–[10], the authors have tried to determine the attribute that should be measured to assess the performance of EMS.

Many models exist in literature like Location Set Covering Problem (LSCP), Maximal Covering Location Problem (MCLP), Maximum Expected Covering Location Problem (MEXCLP), and Double Standard Model (DSM) for solving the facility location problem that has considered area covered by EMS as a significant attribute for proposing allocation plans. Similarly, some papers in the literature have considered service distance as an attribute for evaluating the performance of EMSs. In addition, the background of many articles has validated response time as a significant factor for gaining better insight into the operational performance of EMS. Some works have demonstrated that the patient's mortality and recovery rates are highly influenced by the response time [11]–[14]. Using the same objective in this paper, the authors have considered response time as the prime attribute for estimating and improving the medical service provided by EMS.

EMS operates in an uncertain environment in terms of demand rate, travel time, and response time. The effect of such an uncertain environment on the working of EMS was studied using a simulation-based evaluation method by Ünlüyurt and Tunçer [15]. Different authors have used various algorithms to attain optimized ambulance allocation plans. A data-driven approach was used to allocate and dynamically relocate the complete fleet of ambulances, using a mixed-integer linear formulation to solve the problem of ambulance allocation [16]. Firooze et al. proposed an optimizing model for allocating the ambulances to the base stations, considering the capacity of every base station, travel time, and service time of ambulances [17]. The authors [10], [18] proposed a new model by integrating survival functions with a motive to capture different categories of patients. The conditions faced by the EMS organization are dynamic in terms of variation in travel time, frequency of requests, speed of ambulances, and coverage of areas while fulfilling the requests. The critical issue of the impact of spatial randomness of demand has been considered in very few studies failing to obtain appropriate solutions. The covering models may not be suitable because the spatial distribution of demands may or may not be covered entirely, violating the assumption of the all-or-nothing binary coverage [19]. In 2018, an adjusted queuing solution for ambulance allocation was proposed by considering a heterogeneous spatial distribution of demands in urban and rural areas [20]. Although this solution helped overcome the overstaffing problem, it ignored the definite spatial distribution of demand in each area. In another solution by Chen et al., the authors used various shapes and sizes of grid systems to overcome the problem of the spatial distribution of demand [21]. However, obtaining a probability density function for request calls in a specific grid is challenging [22]. Moreover, a grid area holds no importance until it is classified as an area with a high frequency of request calls, a hospital, or a community resulting in an unstable demand distribution in the grid.

Considering these factors, Degel et al. proposed a time-dependent ambulance allocation model to improve the quality of emergency services [23]. In another work, the authors used a robust optimization approach to improve the functioning of EMS, considering the spatial demand characteristics and uncertain travel time to the requested site [24]. Geroliminis et al. presented a model and a heuristic solution for the optimal deployment of ambulances. They integrated a location model, Genetic Algorithm (GA), and hypercube model for their work [25]. The work by [26] used GA to propose an optimized solution for ambulance deployment. Later, a simulation model incorporating a Gaussian-process-based search algorithm was proposed to attain an optimal allocation plan for ambulances [27]. The authors in [28] handled the emergency department's overcrowding problem by using game theory-based optimization to propose a new optimized allocation plan for ambulances to reduce patient waiting and treatment time. Similarly, many authors [29]–[31] have used Particle Swarm Optimization (PSO) algorithm to achieve an optimized ambulance allocation plan for ambulances. Work was proposed by the authors [32], where a solution for optimally allocating the ambulances was proposed using Jumping PSO. Ant Colony Optimization (ACO) was also used for deploying and redeploying ambulances by [33], [34]. Another algorithm called Shuffled Frog Leaping Algorithm (SFLA) has been used in some works to explore the field of EMS. The authors [35], [36] used SFLA to optimize the working of EMS. SFLA has also been used in different domains for optimally allocating resources [37], [38]. However, there is a dearth of research papers where SFLA is used in the context of EMS.

Research by Elbeltagi et al. [38] indicated that SFLA is a better optimization technique than other evolutionary algorithms such as PSO, ACO, GA, and memetic algorithms. SFLA is a memetic metaheuristic algorithm proposed by Eusuff and Lansey in 2003. This algorithm combines the benefit of the social behavior-based PSO algorithm and genetic-based memetic algorithm. It performs similarly to PSO and surpasses GA in terms of quality of solution, consistency, and processing time. It is considered an efficient and fast algorithm as it can quickly converge to global optima with small population size. However, in some cases, traditional SFLA traps in local optima. To avoid this issue, the authors have proposed mutation-based SFLA (mSFLA) by incorporating the concept of mutation into the working of SFLA.

Considering the previous studies focusing on improvising EMS, most works have used simplified assumptions to come up with a result, while others fail to provide a mutual comparison of models. Computer simulation has proved to be the best way to assess the validation of different processes. Due to the lack of operational data or to avoid ample computation time, many simulation models oversimplify the actual operation. However, the simulation model should consider all the processes, sub-processes, and real-time conditions to find accurate results. The actual operation of the system should be considered while deriving the parameters for the simulation model. Therefore, in this work, the authors have proposed and used a simulation-optimization framework with mSFLA as the optimization component for ambulance allocation. The work



considers the spatial distribution of demands and other uncertainty factors associated with the environment of EMS. To verify and validate the suitability of mSFLA, the algorithm is executed in the MATLAB environment to compare the results with PSO and GA.

### III. PROBLEM BACKGROUND

In India, EMS refers to the ambulance service provided for on-the-spot treatment by paramedics or transporting sick or accident victims to the hospital. Despite being an essential component of society, a fully encompassing definition of EMS is impossible as it does not have a strong representation at the federal level owing to numerous local agencies providing EMS to the public. EMS agencies are classified into three categories based on the tasks they perform.

- 1) EMS that handles scheduled medical transport,
- 2) EMS that handles emergent inter-facility transport, and
- 3) EMS agencies that primarily handle 102-based emergency calls with or without transport.

In this study, the authors have focused on the third category that deals mainly with the optimized use of ambulances. In terms of population and vehicle density across any country, extensive growth is visible. With an increase in vehicle density, accidents (fatal and non-fatal) have also increased, thus, raising the concern of providing medical facilities at the location of the accident. Therefore, when an accident occurs at any location, an ambulance or hospital should be within reach in the shortest duration possible. Since setting up hospitals in every area is impossible, ambulances can be strategically deployed so that on-the-spot treatment and transportation can be provided to accident victims at the earliest. Centralized Accident and Trauma Services (CATS) is an autonomous body of government in Delhi, India, that provides EMS to the victims of accidents and trauma with an ART of approximately 13 minutes. CATS has deployed 50 ambulances at 11 base stations covering the southern portion of Delhi. The area of Southern Delhi is approximately 857 square kilometers and comprises four districts South West Delhi, South East Delhi, New Delhi, and South Delhi. The high frequency of request calls from Southern Delhi motivated the authors to select and work upon the data of this region to allocate and dispatch ambulances for handling accidents and reducing the casualty rate.

### IV. PROBLEM FORMULATION

The problem of allocating ambulances involves distributing a specific count of ambulances ( $A$ ) in the fleet among the base stations ( $B$ ) in such a way that the performance of EMS in terms of response time is improved while serving the requests generated from numerous demand points ( $D$ ). The solution for the distribution of ambulances among the base stations is represented by an integer variable  $a_i$ , where  $i \in B$  specifies the exact count of ambulances placed at different base stations. Thus,

$$A = \{a_1, a_2, a_3, \dots, a_B\}$$

Considering the real-world scenario, the authors have assumed that at an instant, only ' $p$ ' ambulances are available out of ' $A$ ' ambulances to handle the requests as the other

ambulances are busy handling the patients or are on their way back to the base station. The number of ambulances available at each base station ' $i$ ' is denoted as  $a_i(p)$ . To indicate the availability or non-availability of an ambulance at the base station ' $i$ ', a binary variable  $x_i(p)$  is used. Values 1 and 0 for  $x_i(p)$  indicate availability or non-availability at a particular base station. The mathematical formulation for minimizing ART is as follows.

$$\min RT = \sum_{i \in D} v_i * t_{ij} \quad (1)$$

subject to the constraints.

$$\sum_{i \in D} a_i(p) = p \quad (2)$$

$$\sum a_i = A \quad (3)$$

$$x_i(p) \leq a_i(p) \quad (4)$$

$$a_i(p) \geq 0 \quad (5) \quad x_i(p) \geq 0 \quad (6)$$

$$x_i(p) \in (0,1) \quad (7)$$

In the objective function shown in Equation 1,  $v_i$  indicates the arrival rate of request calls per hour and  $t_{ij}$  denotes the travel time from location  $i$  (base station) to location  $j$  (demand spot). As stated if ' $p$ ' ambulances are available out of  $A$  ambulances, Constraint (2) checks that the total ambulances available at each base station are equal to the total number of ambulances present in the system at the same instant. Constraint (3) restricts the total count of ambulances to  $A$ . The fulfillment of requests is constrained by the presence of a certain number of ambulances at the base station by Constraints (4)-(6). Constraint (7) restricts the value of the variable.

### V. SIMULATION MODELLING

The essential processes of an EMS system are structured in a simulation model, as shown in Fig. 1. As soon as the EMS center receives a call after an emergency occurs, the dispatcher selects the nearest ambulance available to serve at the accident site. The ambulance commutes from the base station to the requested site to locate the victim and provides on-the-spot treatment. After assessing the victim's situation, an ambulance transports the victim to the hospital if needed. The ambulance crew transfers the victim to the emergency room at the hospital. After completing the call, either at the requested site or the hospital, the ambulance returns to the base station and waits until assigned a new task.

An EMS system's complicated structure and process dynamics can be effectively represented with the help of an operational model. Fig. 2 shows the operational model designed using the workflow of EMS to achieve the objective of the proposed work. The emergency calls exhibit strong randomness in dimensions of time and space as they can occur at any time and place. Therefore, the spatiotemporal randomness of such calls should be described quantitatively for the correct working of the simulation model. Hence, all the data used in the simulation should be defined accurately to attain relevant results. A random two-dimensional variable ' $r$ ' represents the latitude and longitude coordinates of the point from where the request initiates. The EMS system uses these coordinates to know the exact request location to respond to

emergency calls. The optimization algorithm used in work determines the state values for base stations and ambulances. These state values act as input for the simulation model. The state values refer to the data on the location (coordinates) of the base station, the id of the base station, and the number of ambulances available at that base station. Whenever a request call arrives, the travel time between the request location 'r' and each base station 'bs<sub>i</sub>' can be calculated using the coordinate

data of 'r', base stations, and Google distance matrix application programming interface. This way, a list of base stations ranked by travel time from the requested location is obtained. Then an ambulance with the status 'available' is selected from the base station with the least travel time to request location and dispatched to serve the patient. The status of the selected ambulance is changed from 'available' to 'busy.'

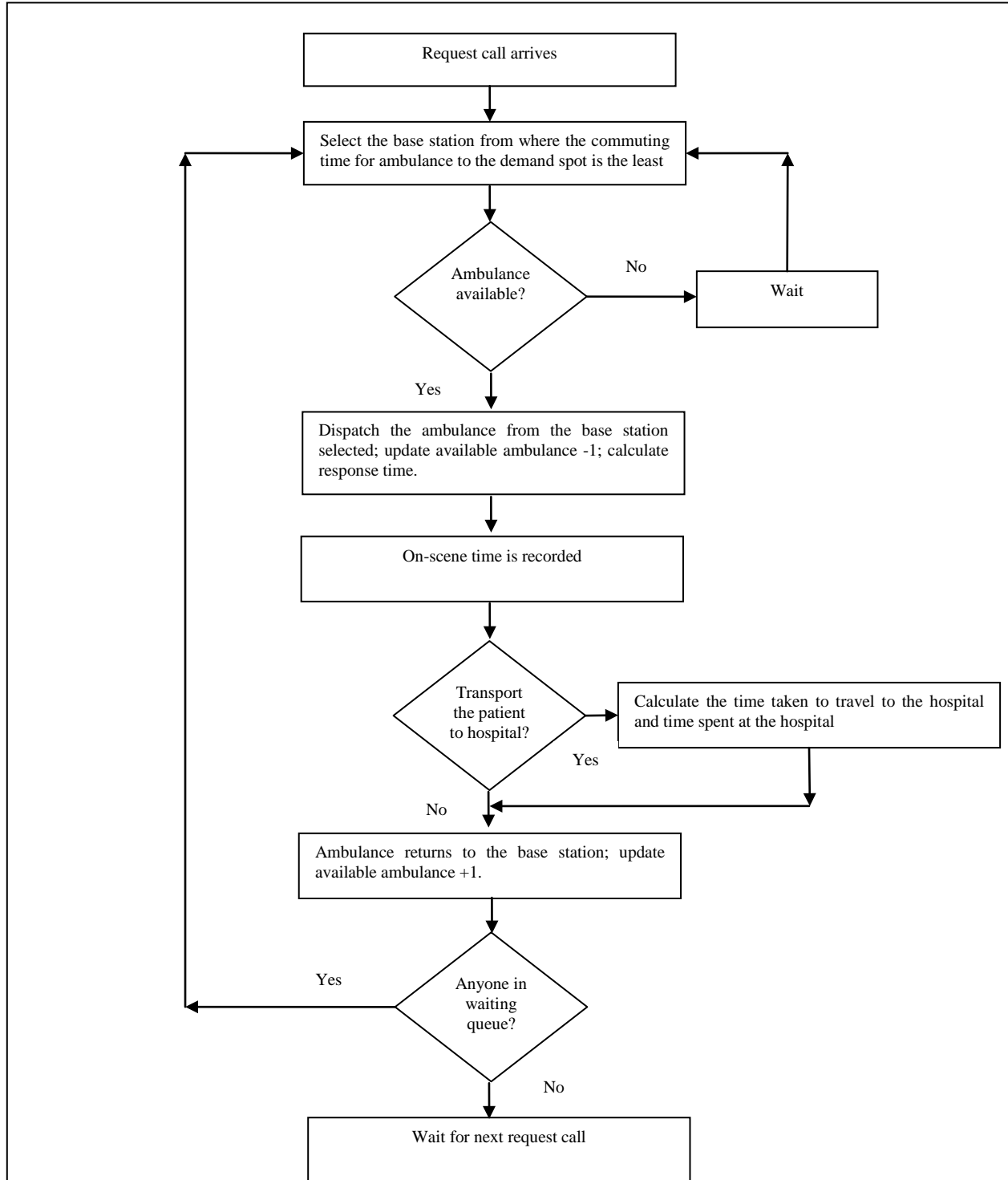


Fig. 1. Simulation Model of EMS System.

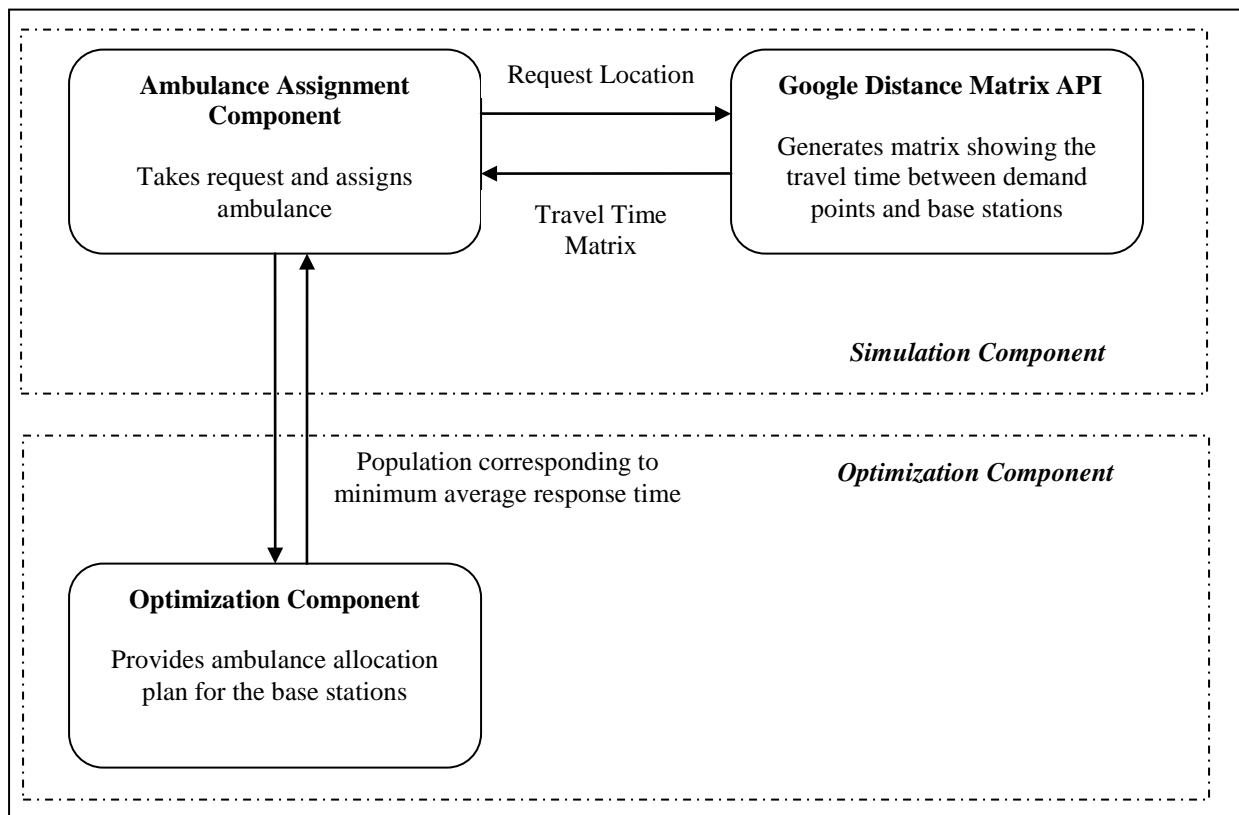


Fig. 2. Operational Model for the Proposed Work.

After the ambulance dispatches, the system calculates response and delay time values. Response time is the time an ambulance takes to reach the patient's location after the request is initiated. As soon as the ambulance arrives at the requested site, the simulation model records the time for on-the-spot treatment. If the patient does not require transportation to a hospital, the ambulance returns to the base station, and its state changes from busy to available. Otherwise, travel and waiting time values are calculated if the patient is transported to a hospital. If no ambulance is available, the patient has to wait in a queue and is served as soon as an ambulance is available. The response time, in this case, is calculated as the sum of travel time and waiting time. After all the requests are served in a day, the total time taken by ambulances to reach the requested locations (response time) is divided by the total number of request calls received at the EMS centre to find the value of ART. The algorithm of mSFLA has been used as an optimizing component and works on the result of the simulation component to find the best allocation plan for ambulances.

#### A. Shuffled Frog Leaping Algorithm

The shuffled frog leaping algorithm helps in finding an optimal solution. It is a memetic meta-heuristic population-based cooperative search approach that imitates the group behavior (jumping strategy) used by frogs to find a location with the maximum amount of food. The algorithm incorporates techniques for performing local searches and exchanging global information. The frogs are randomly assigned a location in the search space. Several groups are formed by dividing the population of frogs, thus generating memeplexes. The memeplexes then evolve separately in different directions

within the search space. Individual frogs can use the information of the population's best frog (global best) or best frog in the memeplex (local best) and change their direction. Each frog experiences memetic evolution because they influence each other and improve their performance to achieve the goal. After a specific number of memetic evolutions, the memeplexes are shuffled to generate new memeplexes, enhancing frogs' ability to attain the best solution in search space. Thus, PSO and Shuffled complex evolution [39] are used for local search and integrating information from parallel local searches in SFLA. The various steps in SFLA are explained below.

Step 1: Population initialization and parameter setting: Various parameters need to be set up for SFLA, such as the size of the population, number of memeplexes and sub memeplexes, and number of memetic evolutions. A random population of 'N' frogs is generated to form the population. For all the frogs in the population, the fitness value is calculated.

Step 2: Grouping: The fitness value obtained above is used to sort the frogs in descending order of their fitness value. These frogs are then divided into 'm' subgroups, with n frogs in each subgroup.

Step 3: Intra-group search: From each subgroup, the frog with the best fitness value 'X<sub>b</sub>' and worst fitness value 'X<sub>w</sub>' are found. The worst solution in the subgroup is updated by using Equation 8 and 9, and only one solution which is the worst in the subgroup is updated at a sub iteration. For updating the position of the worst frog the following equations are used

$$S_i = rand * (X_b - X_w) \quad (8)$$

$$X'_w = X_w + S_i \quad (9)$$

so that

$$S_{imin} < S_i < S_{imax}$$

where  $S_i$  is the variation in the location of frog attained in a single jump. 'rand' is a uniformly distributed random number ranging between 0 and 1. The minimum and maximum step sizes for frogs are represented by  $S_{imin}$  and  $S_{imax}$ . The new position of the worst frog is represented by  $X'_w$ . If the value of  $X'_w$  is better than  $X_w$  then the value of  $X'_w$  replaces the value of  $X_w$  else the new value for  $X'_w$  is calculated by Equation 10 and 11.

$$S_i = rand * (X_{bg} - X_w) \quad (10)$$

$$X'_w = X_w + S_i \quad (11)$$

$X_{bg}$  is the best frog in the current population. In case if the value of  $X'_w$  is still not better than  $X_w$ , then the value of  $X'_w$  can be calculated using

$$X'_w = g + rand(1, D) \otimes (h - g) \quad (12)$$

In the above equation  $D$  represents the dimension of the optimization problem.  $rand(1, D)$  is a random vector of  $D$  components with each component between 0 and 1. 'g' and 'h' represent the upper and lower boundary vectors of the decisive variables.  $\otimes$  means an entry wise multiplication.

The worst frog and best frog is determined from the subgroups attained. Repeated subgroup search is carried out for predefined number of sub iterations. The intra group search stops when the search has been finished by all the subgroups.

Step 4: Exchange of global information: The exchange of global information considers reorganizing all the subgroups into a population of N frogs. Steps 2 and 3 are used again to sort and divide the population into subgroups. Alternate executions of these steps are carried out until either the termination criterion is reached or the best solution is obtained. However, in some cases, a low diversity in the population traps the SFLA into local optima or premature convergence. Therefore, the concept of strong mutation is proposed in this paper to increase the diversity in the population. This concept works upon generating a trial mutated vector using the values of the best solution ( $X_b$ ) in each memplex and the value of the globally best solution ( $X_{bg}$ ). It is crucial to ensure that the dimension of the mutation vector and the number of memplexes is the same.

$$X_{mut}^i = X_{rand}^i + ra(X_b^i - X_{rand}^i) + ra(X_{bg}^i - X_{rand}^i) \quad (13)$$

$i = 1, 2, 3, \dots$  number of memplexes.

Here  $X_{rand}^i$  represents a randomly generated vector and  $ra$  is random number between 0 and 1. Now, the generation cost of trial vector  $f(X_{mut}^i)$  and target vector  $f(X_{bg}^i)$  are compared. If the value of the former is better than the latter, then the target vector is replaced by the trial vector in the next generation. Thus using this step, the algorithm can be prevented from being stuck into a local optimum, and convergence of the algorithm to a global value can be assured.

## B. Application of mSFLA to the Problem of Ambulance Allocation

The application of mSFLA to ambulance allocation problem is explained in this section. The steps used in the work are as follows:

- 1) The coordinate data about the base stations, count of ambulances at the base stations, total number of ambulances in the fleet, coordinate data of demand points are defined.
- 2) Initial population is generated as

$$\text{Population} = \begin{bmatrix} A_1 \\ A_2 \\ A_3 \\ \dots \\ A_{100} \end{bmatrix}$$

$$A_i = a_{i,1}, a_{i,2}, \dots, a_{i,N}$$

- 3) The objective function is defined stating the constraints and values.
- 4) Compute the fitness value for the objective function.
- 5) Sort and divide the population into memplexes on the basis of the value of the fitness function. The local best solution ( $X_b$ ) and the global best solution ( $X_{bg}$ ) is defined.
- 6) The frog with worst solution ( $X_w$ ) is amended using Equation (8) and (9) or Equation (10) and (11) depending on the situation explained in previous section.
- 7) The process of mutation is applied and the result obtained is compared with the value of  $X_{bg}$ .
- 8) The values are updated and the amended and steps through 4 to 8 are repeated until the termination criteria is met.
- 9) The best solution is obtained.

## VI. SIMULATION RESULTS AND DISCUSSION

### A. Area of Concern

The authors have undertaken the southern portion of Delhi to obtain an optimal allocation plan for a fleet of 50 ambulances at 11 base stations in the area. Fig. 3 and Fig. 4 show the southern portion of Delhi and base stations (BS1-BS11) of that area. The traffic police department of Delhi maintains records and releases a report every year stating the details like count of accidents, locations of accidents, time at which the accident took place, information regarding vehicles involved in accidents, etc. To get an insight into the situation, the authors used the report for the year 2019-2020. The report mentions different locations in Delhi which are accident-prone and must be taken into consideration by EMS organizations while devising and designing any policy or strategy. However, since an accident can occur at any location and at any point of time, the authors invented a robust framework that can handle any request initiated from any point in a short time. Considering each point (latitude and longitude) as a tentative spot for the occurrence of an accident, the authors divided southern Delhi into 230 blocks. Each block covered an area of approximately four kilometers square shown in Fig. 5. In every run of the framework, random requests are generated ensuring at least one request is initiated from each block.



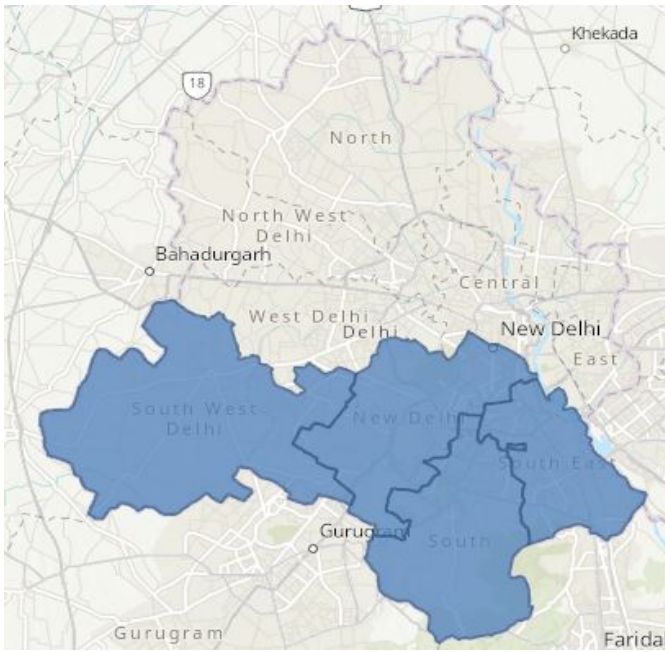


Fig. 3. Southern Portion of Delhi.

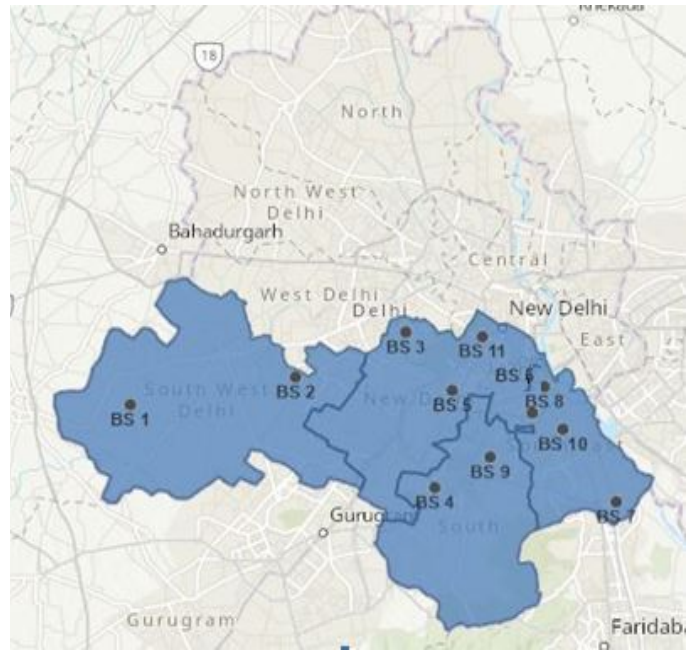


Fig. 4. Base Stations of Southern Delhi.

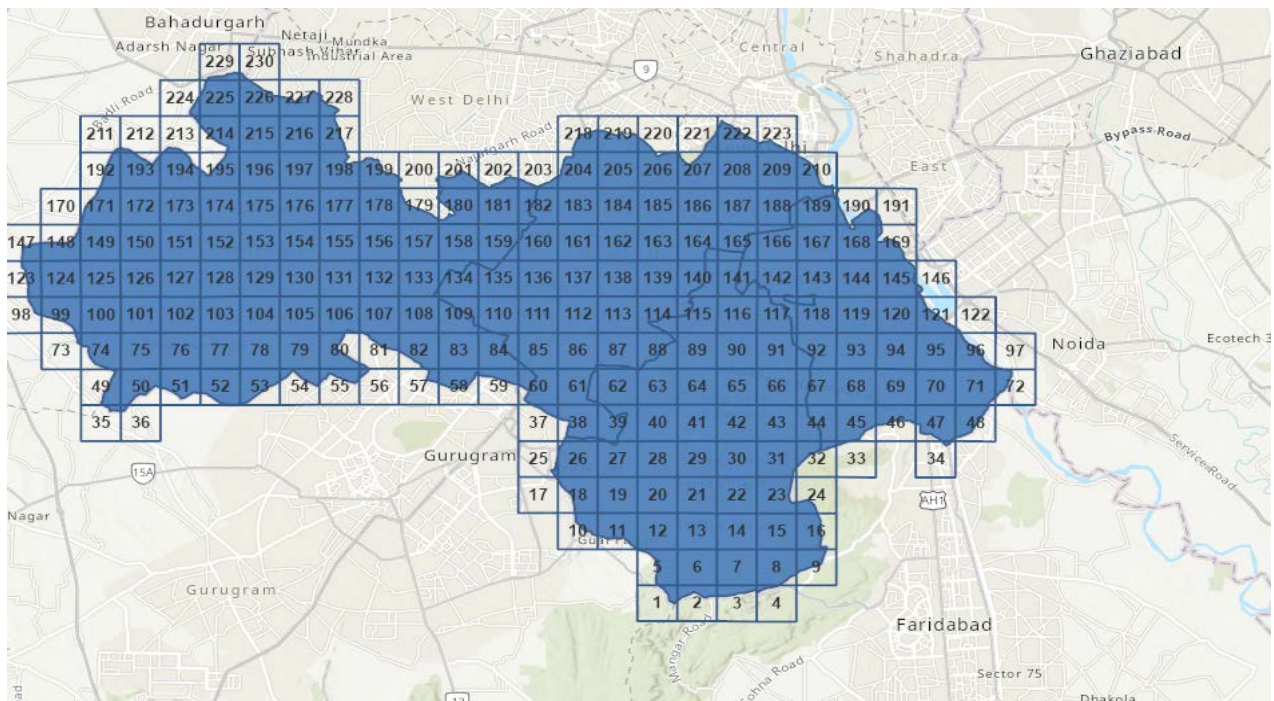


Fig. 5. Tessellations of Southern Delhi.

**B. Results and Discussion**

MATLAB environment is used to execute the work for 15000 requests and a population size of 100. To exemplify the robustness and efficiency of mSFLA, the algorithm is executed as the optimization component in the operational model of the work. For implementing the mSFLA, the authors conducted many experiments to find appropriate values for various parameters. From the result of the experiments, the authors set values for the parameters: number of memplexes, number of frogs in each memplex, iteration count for global exploration,

and iteration count for local exploration, as shown in Table I. The parameter values used for GA and PSO are shown in Table II and Table III respectively.

For comparison, similar simulations are performed using PSO and GA as optimization components in the operational model stated in Section V. 20 runs of simulations are performed with each algorithm to compare the performance of the algorithms using the metrics such as the value of the objective function, convergence rate, and constancy repeatability.

TABLE I. PARAMETERS FOR M SFLA OPTIMIZATION

Parameter Name	Value
Number of frogs in each memplex	10
Number of memplexes	10
Iteration max <sub>1</sub>	70
Iteration max <sub>2</sub>	100
Iteration of mutation	10

TABLE II. PARAMETERS FOR GA

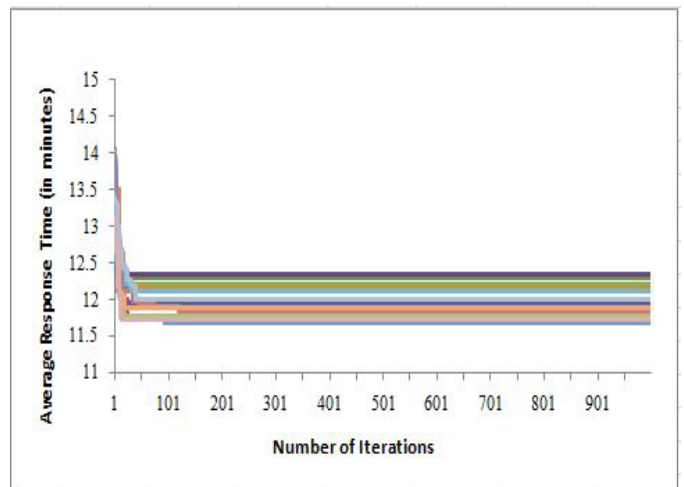
Parameter Name	Value
Population size	100
Number of iterations	1000
Crossover fraction	0.1
Mutation fraction	0.8

TABLE III. PARAMETERS FOR PSO

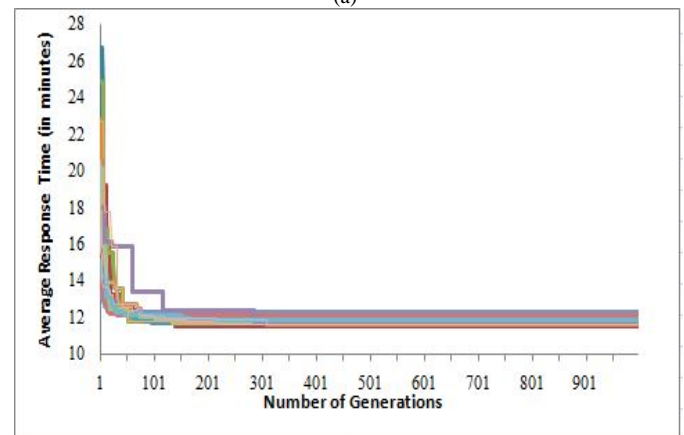
Parameter Name	Value
Population size	100
Number of iterations	1000
Cognitive coefficient (c <sub>1</sub> )	2
Social coefficient (c <sub>2</sub> )	2
Inertia coefficient (w)	0.8

1) *Convergence rate*: The suitability of an algorithm for an optimization problem can be evaluated using convergence rate [36]. The convergence graph can also estimate an algorithm's best, average, worst results, and standard deviation. The convergence graph for PSO, GA, and mSFLA is shown in Fig. 6(a), 6(b), and 6(c). As stated, 20 different runs were carried out for all three algorithms to attain the global fitness value for the objective function. The global fitness value is the best fitness value obtained in each iteration within the defined population size of the algorithm. The graph illustrates that the value of global fitness (in the best iteration/generation of every algorithm) changed from 13.58172 to 11.9733 in the case of PSO, 19.52089 to 11.72735 in the case of GA, and 12.93752 to 11.4127 in the case of mSFLA in 1000 iterations of each. The graph also indicates that the PSO algorithm converged in 94, the GA in 358, and the mSFLA in 41 iterations. The quick convergence of mSFLA indicates that the convergence rate and computational time of mSFLA are better than GA and PSO.

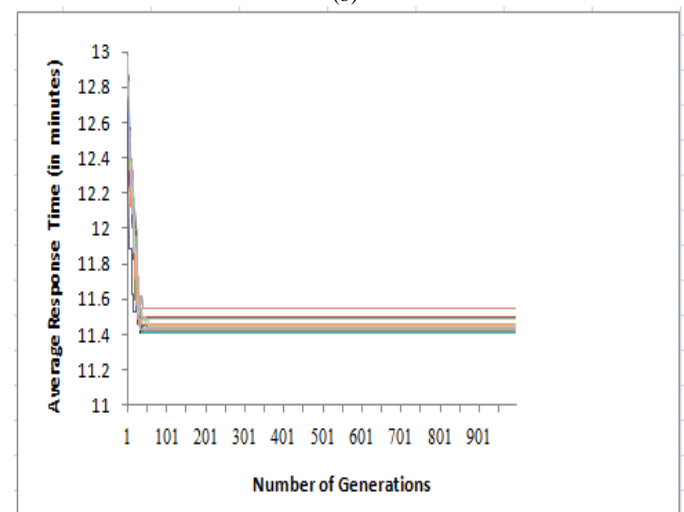
Another comparison result is shown in Fig. 7 and Fig. 8, where the values of standard deviation, best, average, and worst solutions are plotted for each algorithm. The result in Fig. 7 indicates that the even worst solution (maximum ART) provided by mSFLA is better than the other two algorithms' average solutions. In addition, the average result of mSFLA is also better, making it more efficient. The value of Standard Deviation (stdev) for all the algorithms shown in Fig. 8 reveals that the value of stdev i.e. 0.045331 is almost negligible in the case of mSFLA; therefore, it can provide the best solution in each run.



(a)



(b)



(c)

Fig. 6. (a). Convergence Graph of PSO, (b). Convergence Graph of GA, (c). Convergence Graph of mSFLA.



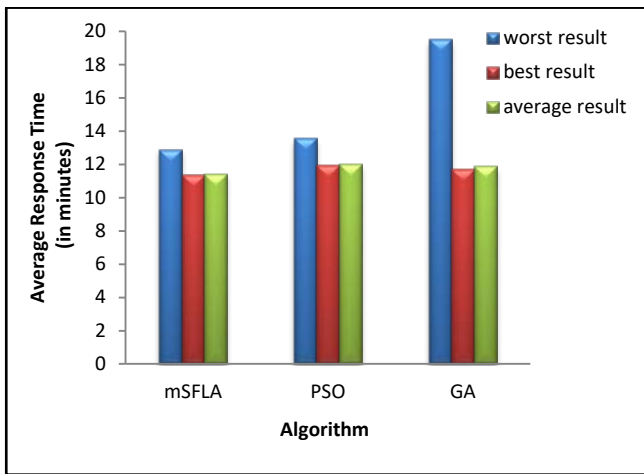


Fig. 7. Comparison of Worst, Best, and Average Result.

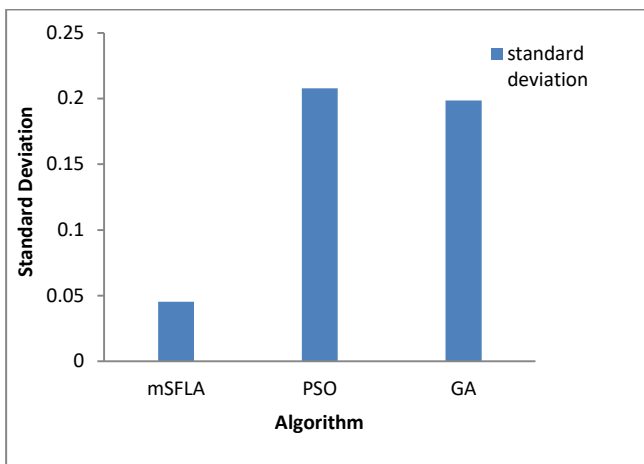


Fig. 8. Comparison of Standard Deviation.

2) *Objective function value:* In this work, the objective function minimizes the value of ART of EMS to provide a prompt service to the people in need. The evolution graph in Fig. 9 depicts the global fitness values of ART for PSO, GA, and mSFLA are 11.9733 minutes, 11.72735 minutes, and 11.4127 minutes respectively.

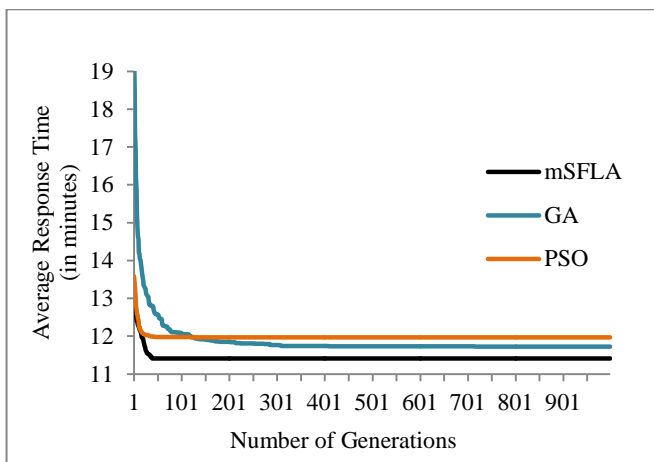
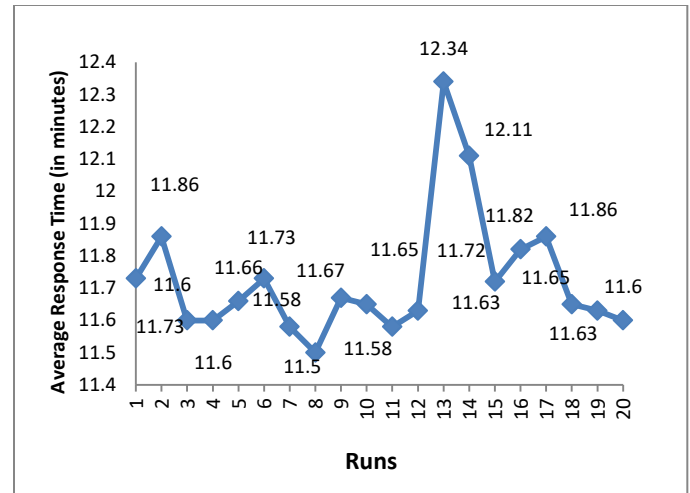
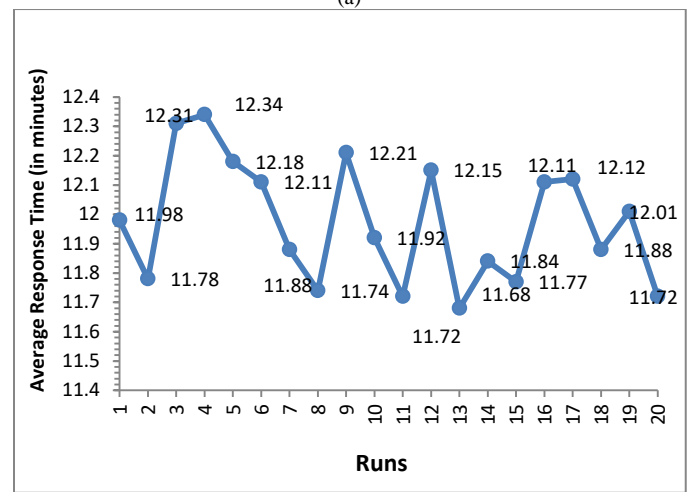


Fig. 9. Evolution Graph of Algorithms.

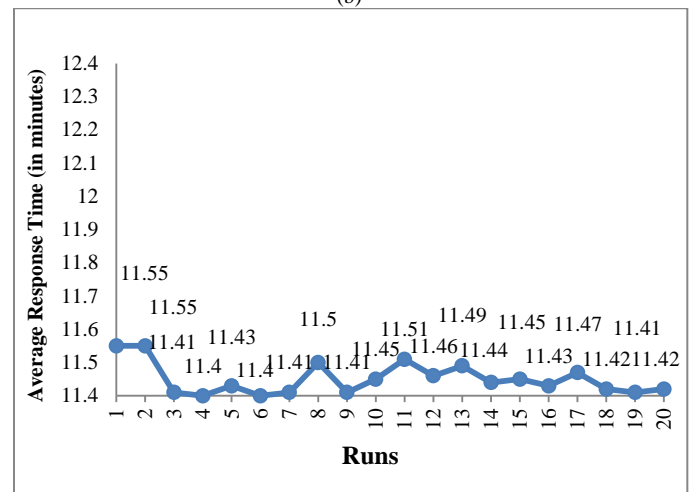
The convergence of mSFLA to the least value demonstrates that the result and performance of mSFLA is better than PSO and GA for the problem of ambulance allocation.



(a)



(b)



(c)

Fig. 10. (a). Constancy Graph of GA, (b). Constancy Graph of PSO, (c). Constancy Graph of mSFLA.

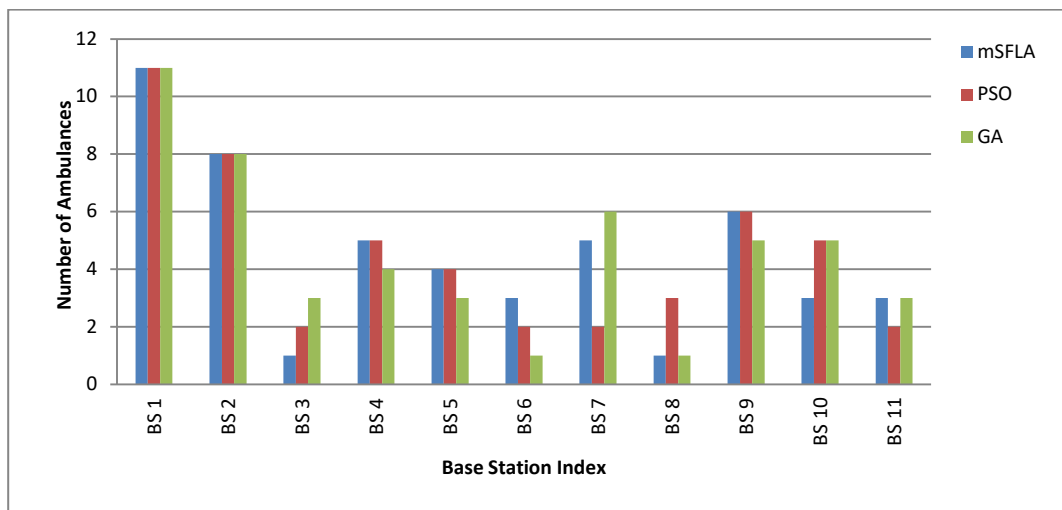


Fig. 11. Ambulance Allocation Plan Attained by Each Algorithm.

3) Constancy repeatability: The performance of any algorithm can also be measured using the concept of constancy repeatability. Constancy repeatability of an algorithm is the similarity rate of the results obtained by the algorithm in different executions with the same input values [40]. To infer the constancy and repeatability of the algorithm, the authors plotted the fitness values obtained by all three algorithms in twenty runs of the operational model. Fig. 10(a), 10(b), and 10(c) show that the changes in the fitness value of GA, PSO, and mSFLA. The changes in the value of GA are from 12.34 to 11.50 minutes, in PSO are from 12.34 to 11.68 minutes, and in mSFLA are from 11.55 to 11.41 minutes. To be more exact, the variance of the results is calculated. The variance values are 0.043178 for PSO, 0.038826 for GA, and 0.00205 for mSFLA. The consistency of any algorithm can be highlighted with the value of variance ranging between 0 and 1. In the proposed work, the variance value is between 0 and 1 for all three algorithms stating that all the algorithms are stable and consistent. However, the global optima results obtained by mSFLA are close to the average value in most runs, so it characterizes mSFLA as the most stable algorithm among the three algorithms. In other words, it can be said that in most cases, mSFLA will converge to global optima or near global optima. The ambulance allocation plan for the area of Southern Delhi provided by each algorithm is shown in Fig. 11.

## VII. CONCLUSION

The performance of EMS significantly affects a country's healthcare system as it is considered responsible for saving people's lives. Response time is considered a key indicator to measure the performance of EMS by evaluating the time an ambulance takes to report at the spot from where the request was generated. To reduce the response time of EMS, the ambulances should be strategically allocated at the base stations so that the commuting time of the ambulance from the base station to the demanded spot is reduced. Considering this motive, the authors undertook the problem of finding an optimal allocation plan for a fleet of 50 ambulances among the

11 base stations in the southern portion of Delhi. The authors used an operational model that showed the flow of data between the simulation component and optimization component. For the optimization component, the authors proposed mSFLA that used the concept of mutation in SFLA to avoid being trapped in local optima. mSFLA was compared with GA and PSO using different metrics. The results shown in Section VI help analyze the performance of mSFLA with PSO and GA. The objective of the work to attain an allocation plan with minimum response time is attained by mSFLA. It is able to reduce the value of ART from 13 minutes to 11.41 minutes, i.e., by 12.23%.

A comparison of standard deviation, best and worst solutions of the proposed algorithm proves that mSFLA is more effective than the other two algorithms. The small value of 0.045331 for standard deviation signifies that mSFLA is consistent and reliable. Quick convergence and short execution time of mSFLA imply that it can be efficiently utilized in optimization problems similar to ambulance allocation problems. Moreover, mSFLA converges to a global optima value of 11.4127 minutes at lower iterations i.e. 41<sup>st</sup> iteration number taking less execution time than PSO and GA, which converge to a global optima value of 11.9733 and 11.72735 in 94<sup>th</sup> and 358<sup>th</sup> iteration number at much higher iterations. Therefore, mSFLA appears superior to the other two algorithms regarding the quality of solution and convergence rapidity. This work also validates the competency of mSFLA to other algorithms in handling problems similar to allocation problems.

As it is impossible to consider all the possible scenarios, the authors would like to extend the work by changing the single objective function to a multiobjective function. In addition, the authors will also focus on proposing an efficient strategy and solution for dynamically allocating and relocating ambulances.

## REFERENCES

- [1] V. Bélanger, A. Ruiz, and P. Soriano, "Recent optimization models and trends in location, relocation, and dispatching of emergency medical vehicles," *Eur. J. Oper. Res.*, vol. 272, no. 1, pp. 1–23, 2019.
- [2] H. Andersson, T. A. Granberg, M. Christiansen, E. S. Aartun, and H. Leknes, "Using optimization to provide decision support for strategic

- emergency medical service planning – Three case studies,” *Int. J. Med. Inform.*, vol. 133, no. July 2019, pp. 103–113, 2020.
- [3] T. Andersson and P. Värbrand, “Decision support tools for ambulance dispatch and relocation,” *J. Oper. Res. Soc.*, vol. 58, no. 2, pp. 195–201, 2007.
- [4] C. J. Jagtenberg, S. Bhulai, and R. D. van der Mei, “Dynamic ambulance dispatching: is the closest-idle policy always optimal?,” *Health Care Manag. Sci.*, vol. 20, no. 4, pp. 517–531, 2017.
- [5] S. El-Masri and B. Saddik, “An emergency system to improve ambulance dispatching, ambulance diversion and clinical handover communication: a proposed model,” *J. Med. Syst.*, vol. 36, no. 6, pp. 3917–3923, 2012.
- [6] Y. A. Kochetov and N. B. Shamray, “Optimization of the Ambulance Fleet Location and Relocation,” *J. Appl. Ind. Math.*, vol. 15, pp. 234–252, 2021.
- [7] M. Reuter-Oppermann, P. L. van den Berg, and J. L. Vile, “Logistics for Emergency Medical Service systems,” *Heal. Syst.*, vol. 6, no. 3, pp. 187–208, 2017.
- [8] L. Yan, P. Wang, J. Yang, Y. Hu, Y. Han, and J. Yao, “Refined Path Planning for Emergency Rescue Vehicles on Congested Urban Arterial Roads via Reinforcement Learning Approach,” *J. Adv. Transp.*, vol. 2021, pp. 1–12, 2021.
- [9] M. Eusuff, K. Lansey, and F. Pasha, “Shuffled frog-leaping algorithm: a memetic meta-heuristic for discrete optimization,” *Eng. Optim.*, vol. 38, no. 2, pp. 129–154, 2006.
- [10] V. A. Knight, P. R. Harper, and L. Smith, “Ambulance allocation for maximal survival with heterogeneous outcome measures,” *Omega*, vol. 40, no. 6, pp. 918–926, 2012.
- [11] E. Wilde, “Do emergency medical system response times matter for health outcomes?,” *Health Econ.*, vol. 22, no. 7, pp. 790–806, 2013.
- [12] M. A. Zaffar, H. K. Rajagopalan, C. Saydam, M. Mayorga, and E. Sharer, “Coverage, survivability or response time: A comparative study of performance statistics used in ambulance location models via simulation--optimization,” *Oper. Res. Heal. Care*, vol. 11, pp. 1–12, 2016.
- [13] J. J. Boutillier and T. C. Y. Chan, “Ambulance emergency response optimization in developing countries,” *Oper. Res.*, vol. 68, no. 5, pp. 1315–1334, 2020.
- [14] E. Jánošíková, P. Jankovič, M. Kvet, and F. Zajacová, “Coverage versus response time objectives in ambulance location,” *Int. J. Health Geogr.*, vol. 20, no. 1, pp. 1–16, 2021.
- [15] T. Ünlüyurt and Y. Tunçer, “Estimating the performance of emergency medical service location models via discrete event simulation,” *Comput. Ind. Eng.*, vol. 102, pp. 467–475, 2016.
- [16] S. Saisubramanian, P. Varakantham, and H. C. Lau, “Risk based optimization for improving emergency medical systems,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2015, pp. 17–22.
- [17] S. Firooze, M. Rafiee, and S. M. Zenouzzadeh, “An optimization model for emergency vehicle location and relocation with consideration of unavailability time,” *Sci. Iran.*, vol. 25, no. 6, pp. 3685–3699, 2018.
- [18] R. McCormack and G. Coates, “A simulation model to enable the optimization of ambulance fleet allocation and base station location for increased patient survival,” *Eur. J. Oper. Res.*, vol. 247, no. 1, pp. 294–309, 2015.
- [19] R. Wei, “Coverage location models: alternatives, approximation, and uncertainty,” *Int. Reg. Sci. Rev.*, vol. 39, no. 1, pp. 48–76, 2016.
- [20] M. van Buuren, R. an der Mei, and S. Bhulai, “Demand-point constrained EMS vehicle allocation problems for regions with both urban and rural areas,” *Oper. Res. Heal. Care*, vol. 18, pp. 65–83, 2018.
- [21] A. Y. Chen, T.-Y. Lu, M. H.-M. Ma, and W.-Z. Sun, “Demand Forecast Using Data Analytics for the Preallocation of Ambulances,” *IEEE J. Biomed. Heal. Informatics*, vol. 20, no. 4, pp. 1178–1187, 2015.
- [22] S. Wajid, N. Nezamuddin, and A. Unnikrishnan, “Optimizing Ambulance Locations for Coverage Enhancement of Accident Sites in South Delhi,” *Transp. Res. procedia*, vol. 48, pp. 280–289, 2020.
- [23] D. Degel, L. Wiesche, S. Rachuba, and B. Werners, “Time-dependent ambulance allocation considering data-driven empirically required coverage,” *Health Care Manag. Sci.*, vol. 18, no. 4, pp. 444–458, 2015.
- [24] J. J. Boutillier and T. C. Chan, “Ambulance emergency response optimization in developing countries,” *Oper. Res.*, vol. 68, no. 51, pp. 1315–1334, 2020.
- [25] N. Geroliminis, K. Kepaptsoglou, and M. G. Karlaftis, “A hybrid hypercube--genetic algorithm approach for deploying many emergency response mobile units in an urban network,” *Eur. J. Oper. Res.*, vol. 210, no. 2, pp. 287–300, 2011.
- [26] L. Zhen, K. Wang, H. Hu, and D. Chang, “A simulation optimization framework for ambulance deployment and relocation problems,” *Comput. Ind. Eng.*, vol. 72, pp. 12–23, 2014.
- [27] W. Yang, Q. Su, S. H. Huang, Q. Wang, Y. Zhu, and M. Zhou, “Simulation modeling and optimization for ambulance allocation considering spatiotemporal stochastic demand,” *J. Manag. Sci. Eng.*, vol. 4, no. 4, pp. 252–265, 2019.
- [28] A. JA, J. Zayas-Castro, and H. Charkhgard, “Ambulance allocation optimization model for the overcrowding problem in US emergency departments: A case study in Florida,” *Socio-Economic Plan. Sci.*, vol. 71, p. 100747, 2020.
- [29] AhmadM.Manasrah and H. B. Ali, “Workflow Scheduling Using Hybrid GA-PSO Algorithm in Cloud Computing,” *Wirel. Commun. Mob. Comput.*, vol. 7, no. 4, pp. 17–34, 2018.
- [30] S. H. R. Hajipour, Vahid and Pasandideh, “Proposing an adaptive particle swarm optimization for a novel bi-objective queuing facility location model,” *Econ. Comput. Econ. Cybern. Stud. Res.*, vol. 46, pp. 223–240, 2012.
- [31] H. WA, L. CS, A. AF, A. MH, and T. SS., “Solving maximal covering location with particle swarm optimization,” *Int. J. Eng. Technol.*, vol. 5, no. 4, pp. 3301–3306, 2013.
- [32] Y. Tsai, C. KW, Y. GT, and L. HJ, “Demand forecast and multi-objective ambulance allocation,” *Int. J. Pattern Recognit. Artif. Intell.*, vol. 32, no. 7, p. 1859011, 2018.
- [33] M. Benabdouallah, C. Bojji, and O. El Yaakoubi, “Deployment and redeployment of ambulances using a heuristic method and an ant colony optimization—case study,” in *Third International Conference on Systems of Collaboration (SysCo)*, 2016, pp. 1–4.
- [34] Q. Su, Q. Luo, and H. H. Samuel, “Cost-effective analyses for emergency medical services deployment: A case study in Shanghai,” *Int. J. Prod. Econ.*, vol. 163, pp. 112–123, 2015.
- [35] H. Adarang, A. Bozorgi-Amiri, K. Khalili-Damghani, and R. Tavakkoli-Moghaddam, “A robust bi-objective location-routing model for providing emergency medical services,” *J. Humanit. Logist. Supply Chain Manag.*, 2020.
- [36] X. Duan, T. Niu, and Q. Huang, “An improved shuffled frog leaping algorithm and its application in dynamic emergency vehicle dispatching,” *Math. Probl. Eng.*, pp. 17–28, 2018.
- [37] H. P. Hsu and T. L. Chiang, “An improved shuffled frog-leaping algorithm for solving the dynamic and continuous berth allocation problem (DCBAP),” *Appl. Sci.*, vol. 9, no. 21, p. 4682, 2019.
- [38] Q. Huang and W. Song, “A land-use spatial optimum allocation model coupling a multi-agent system with the shuffled frog leaping algorithm,” *Comput. Environ. Urban Syst.*, vol. 77, p. 101360, 2019.
- [39] Q. Y. Duan, V. K. Gupta, and S. Sorooshian, “Shuffled complex evolution approach for effective and efficient global minimization,” *J. Optim. Theory Appl.*, vol. 76, no. 3, pp. 501–521, 1993.
- [40] B. Saeidian, M. S. Mesgar, B. Pradhan, and M. Ghodousi, “Optimized location-allocation of earthquake relief centers using PSO and ACO, complemented by GIS, clustering, and TOPSIS,” *Int. J. Geo-Information*, vol. 7, no. 8, pp. 292–315, 2018.

# Adopting a Digital Transformation in Moroccan Research Structure using a Knowledge Management System: Case of a Research Laboratory

Fatima-Ezzahra AIT-BENNACER<sup>1</sup>, Abdessadek AAROUD<sup>2</sup>, Khalid AKODADI<sup>3</sup>, Bouchaib CHERRADI<sup>4</sup>  
LaROSERI Laboratory, Faculty of Sciences, Chouaib Doukkali University, El Jadida, 24000, Morocco<sup>1,2,3,4</sup>  
STIE Team, CRMEF Casablanca-Settat, Provincial Section of El Jadida, El Jadida, 24000, Morocco<sup>4</sup>

**Abstract**—Digital Transformation has become one of the most discussed debates; many sectors have adopted digital transformation to gain a competitive advantage and to ensure their continuity. Moroccan universities, in their turn, are facing strategic and managerial challenges due to emerging practices related to digital transformation. To address this issue, the proposed work sets out to define the factors that lead us to adopt a digital transformation using SWOT analysis and to apply total quality management techniques to contribute to our research laboratory's digital transformation, by digitalizing and managing knowledge and processes. KMS-TQM digital platform has been used to capitalize knowledge and profile the different existing functions, positions, tasks, and referential competencies. Then, we analyzed all the actual processes to propose a business process re-engineering using Bizagi Modeler. The study's contribution is to standardize all the current processes in the laboratory to help the Doctoral Studies Center successfully carry out the digital transformation. Moreover, the aim is to make all functions and tasks for each position explicit.

**Keywords**—Business process re-engineering; digital transformation; knowledge management system; Moroccan research laboratory; total quality management

## I. INTRODUCTION

Digital Transformation (DT) has recently increasingly driven organizations to change [1]. It has become one of the most discussed debates in business and organizational contexts [2]. The DT has involved sustainable management in coping with these changes; this is a vital and essential process for organizations that pretend to be leaders of change and be competitive in their sector [3], [4].

COVID-19 pandemic has mobilized research community for developing early diagnosis systems [5]–[11]. In addition, this pandemic has confirmed the need to digitalize public and private organizations (companies and educational institutions, etc.). The pandemic was an opportunity to innovate and accelerate the digital transformation to ensure the continuity and sustainability of the company [12].

In the education sector, the digital transformation has implicated sustainable management in dealing with these transformations [13]–[15]. The DT has been significant as a primary focus for higher education institutions (HEIs) [16].

Many studies have introduced the digital transformation in public administrations and the public sector [17], [18], that are

focusing on redefining their processes to create new forms of public administration and interactions with users of their services.

In the current research, we will focus on the Moroccan research structures, taking as a case study our scientific research laboratory LaROSERI, which unfortunately requires additional efforts to ensure good productivity and sustainability. Through digital transformation, it will be able to manage efficiently and effectively the whole research laboratory from different axes: performance, knowledge management, and business process re-engineering.

In our previous work [19], we considered the research laboratory as a non-profit organization, and then we defined an indicator called “Global Laboratory Performance Indicator GLPI” to measure the laboratory's global performance. As a result, it has been shown a suitable approach should be adopted.

For this purpose, we suggest the following rankings, which show the necessity to rethink our managerial and strategic organization. To define the gaps that lead us to this study, we refer to the Ranking Web of World Research Centers as an initiative of the Cybermetrics Lab, a research group belonging to the Consejo Superior de Investigaciones Científicas (CSIC)<sup>1</sup>. CSIC is one of the leading essential research organizations in Europe. In 2006, CSIC comprised 126 centers and institutes throughout Spain. CSIC is attached to the Ministry of Education, and its primary goal is to promote scientific research to enhance scientific and technological progress.

The following Table I ranks the top ten research centers/labs, according to the 2019 CSIC ranking.

The Table II describes the global and African ranking of Moroccan research centers. As presented, the Moroccan Institute of Scientific and Technical Information - IMIST is ranked 65<sup>th</sup> in Africa and 4094<sup>th</sup> worldwide.

These rankings highlight the importance and the need for a set of actions to reposition the scientific research structures that are facing strategical and managerial challenges. Furthermore, Morocco upholds digital transformation through many initiatives to accelerate its development; we can cite some of them as described in the Table III.

<sup>1</sup> <https://research.webometrics.info/en/>

TABLE I. CSIC RANKING – 2019

Ranking	Institution	Country
1	National Institutes of Health	USA
2	National Aeronautics and Space Administration	USA
3	Centre National de la Recherche Scientifique CNRS	France
4	Max Planck Gesellschaft	Germany
5	Chinese Academy of Science CAS / 中国科学院	China
6	Centers for Disease Control and Prevention	USA
7	US Department of Veterans Affairs	USA
8	Consejo Superior de Investigaciones Cientificas CSIC	Spain
9	National Oceanic and Atmospheric Administration	USA
10	Consiglio Nazionale delle Ricerche CNR	Italy

TABLE II. CSIC MOROCCAN RESEARCH CENTERS/LABS - AFRICAN RANKING - 2019

Global ranking	African ranking	Research Center/Lab
4094 <sup>th</sup>	65 <sup>th</sup>	Moroccan Institute of Scientific and Technical Information - IMIST
4198 <sup>th</sup>	69 <sup>th</sup>	National Center for Scientific and Technical Research - CNRST
4263 <sup>rd</sup>	70 <sup>th</sup>	Royal Institute of Amazigh Culture - IRCA
4529 <sup>th</sup>	77 <sup>th</sup>	Pasteur Institute of Morocco
5139 <sup>th</sup>	90 <sup>th</sup>	Scientific Institute of Rabat

TABLE III. MOROCCAN INITIATIVES

Moroccan Initiative	Description
“Horizon 2020”	In Morocco, the digital transformation has been accelerated, mainly due to the major government initiatives, "Horizon 2020", launched in 2017, and then "Horizon 2025", which have set ambitious goals in terms of e-government and training of young people in innovative technologies.
The National Plan to reform administration (2018-2021)	It is considered a crucial demand to upgrade the administration and the public Service through its restructuring and the reinforcement of its managerial and technical capacities to be qualified to offer good governance, ensure services of the general interest, and provide users with quality services.
The Framework Law 51-17	The Moroccan parliament approved the framework law 51-17 in August 2019, which concerns the government's strategic plan "2015-2030" to strengthen the national education system. The framework law 51-17 will mandate the creation of a national commission to supervise its execution and the overall education system reform in Morocco.

To properly define the factors that led us to this study, we propose a SWOT analysis as described in Table IV.

As shown in the SWOT analysis, it describes what the research laboratory excels, identifies in which it needs to undertake improvements to remain competitive. In addition, the analysis mentions factors that have the potential to harm the research laboratory or that could provide a strategic advantage to it.

TABLE IV. SWOT ANALYSIS

SWOT analysis	
Internal	
Strengths	Weaknesses
<ul style="list-style-type: none"> <li>- The diversification of the theses topics, which emerge from several fields</li> <li>- Developing scientific and technical research in the computer sciences field</li> <li>- Offering different doctoral training and part-time teaching opportunities for PhDs</li> <li>- Collaborations between research teams, such as thesis co-supervision, and co-authors.</li> <li>- A good work atmosphere</li> </ul>	<ul style="list-style-type: none"> <li>- Insufficient dynamics of doctoral studies (doctoral theses duration are too often exceeding the regulatory period, which is three years)</li> <li>- Lack of a Digital workspace that can promote collaborative work</li> <li>- Insufficient resources (offices and materials) for the expected increase in the number of doctoral students enrolled.</li> <li>- Percentage of funded theses</li> <li>- Lack of using monitoring and steering tools (skills management, performance management, process management)</li> <li>- The absence of an internal structure specialized in setting up and monitoring theses, research projects, and research activities.</li> <li>- Lack of training to prepare PhD students for the job market actions (soft skills, coaching, personal development)</li> <li>- The need for a digital strategy</li> <li>- Lack of a quality management system</li> <li>- Lack of a digital tool to communicate research outcomes within the laboratory members</li> <li>- Lack of positions and skills repository</li> <li>- Integration of new information and communication technologies is relatively modest.</li> <li>- Lack of a digital system for monitoring and evaluating research activity</li> <li>- Lack of using efficient governance tools such as the BSC</li> <li>- Need for a management training, people in positions of responsibility</li> </ul>
External	
Opportunities	Threats
<ul style="list-style-type: none"> <li>- Research in collaboration with external laboratories, universities, or institutions.</li> <li>- Research projects in collaboration with CNRST, OCP, UM6P</li> <li>- Communicating research outcomes in different scientific events</li> <li>- Publishing scientific papers in indexed journals</li> </ul>	<ul style="list-style-type: none"> <li>- Absence of a digital platform</li> <li>- Concurrence with other national institutions</li> <li>- The absence of an alumni network</li> <li>- The gap between ambition and resources allocated to research</li> <li>- Weak private R&amp;D and insufficient business/university interactions</li> <li>- Lack of sufficient anticipation of investments in IT/digital infrastructure</li> </ul>

According to the points above, we have decided that adopting a digital transformation could solve this issue, thus digitizing, managing, and eventually repositioning the research laboratory within the university.

This work emphasizes using a knowledge management system (KMS) capable of filling all the gaps mentioned.

To successfully conduct this research, we focused on the following three research questions:

- How can the laboratory develop the digital transformation strategy?
- How clear does the management in the Lab understand the needs for a digital transformation of its organization?
- How does the laboratory apply digital technologies and managerial practices in its processes?

To answer the research questions, the main contributions of this work are the following:

- Defining the factors and limits that have forced us to adopt a digital transformation using the SWOT analysis.
- Using a Knowledge Management System that combines all the TQM aspects to manage the whole structure.
- Standardize all the research structure processes to be expressed explicitly.

The rest of this paper is structured as follows: Section II presents the related works and the problematic. Section III describes the proposed approach. Section IV shows the implementation using KMS-TQM digital platform and Bizagi Modeler. Finally, section V concludes the current work and proposes the perspectives.

## II. RELATED WORKS

Due to new needs and requirements, several sectors have used digitalization to obtain a competitive advantage and ensure continuity [20], [21]. To discuss the state of the art, we propose the following subsections that determine each aspect of our work: Digitalization and Digital transformation, Knowledge Management, Business process Re-engineering, and their applications in the context of HEIs.

### A. Digitalization and Digital Transformation

According to [12], the COVID-19 pandemic has confirmed the need and importance of digitalization in public and private organizations. It was an opportunity to innovate and accelerate the digital transformation to ensure the continuity and sustainability of organizations. To define the term 'Digitalization' and 'Digital Transformation', the following Table V illustrates some definitions proposed in the literature.

As pointed out in [25], digital competence is considered as a set of skills, knowledge, and attitudes necessary to use ICT and digital devices for responsibilities such as information management, and collaboration in an effective, efficient and ethical way. Digital transformation is considered a well-

known topic at the moment, and ideas for digital products, facilities, and media were already widely understood in the 1990s and 2000s [26].

TABLE V. DIGITALIZATION AND DIGITAL TRANSFORMATION DEFINITIONS

Ref	Definition
[22]	Digitalization is a "sustainable company-level transformation via revised or newly created business operations and business models achieved through value-added digitalization initiatives, ultimately resulting in improved profitability."
[23]	The digitalization is "the application of any digital technologies to all human activities, such as personal life, social, economic and political activities."
[24]	Digital transformation is "a process that aims to improve an entity by triggering significant changes to its properties through combinations of information, computing, communication, and connectivity technologies."

"Each organizational transformation implies a cultural transformation", the introduction of new technologies and digitalization has strongly contributed to the organizational and cultural transformation of companies [27]. Digital transformation has caused a significant change in the business, both in its activities, its organization and even in its culture [28]. Digital transformation is the integration of new processes within the company, such as adopting new technologies, tools, and work methods [29]. It impacts the global functioning of companies and transforms working methods and processes, requiring managerial approaches for a long-term vision to remain competitive, efficient, and modern [30].

### B. Total Quality Management

Total quality management is a managerial approach that began in Japanese industry and has received increased attention in the West since the early 1980s[31], [32]. Total quality refers to a company's culture, attitude, and organization that attempts to consistently offer its customers products and services that fulfill their expectations [33].

TQM is a quality management approach whose target is to achieve ideal quality, the entire company should be mobilized and involved, by reducing waste as much as possible and by continuously improving the output elements [28]. Many agree that the TQM movement began in Japan, the term TQM comes from TQC, it was coined by A.V. Freignbaum, 1983 [34], [35]. Organizations that have successfully used the principles of TQM, have integrated the customer and quality into their business strategy [36]. It is the result of the efforts made to develop Quality Management.

Important aspects of TQM encompass quality management leadership and commitment, continuous improvement, rapid response, evidence-based actions, employee involvement, and a TQM culture [37].

In the context of higher education, several quality management models developed for use in industry have been involved in HEIs around the world [43], such as TQM, EFQM, Balanced scorecard, Malcolm Baldrige award, ISO 9000, Business process re-engineering and SERVQUAL. One of the most well-known quality management models that have



been implemented in higher education is Total Quality Management (TQM) as described in Table VI [44].

The study [45] defines the seven TQM factors as follows: Leadership, Strategic planning, Human resource management, Customer orientation, Process management, Information analysis, and Continuous process improvement. Through the implication of TQM concepts, assists organizations in learning strategies to increase productivity. The dimensions of TQM indicate the broad range of features in organizational cultures that promote innovation. Yet, the success of TQM demands an organizational culture based on trust and knowledge sharing [46].

TABLE VI. TQM APPLICATIONS IN HEIS

Ref	Year	Overview
[38]	2018	Through a review of the literature, this research aims to examine the impact of TQM on the organizational performance of Portuguese universities and polytechnic higher education institutions. The purpose of this study is to point out the importance of quality in education, specifically in HEIs, as indicated by the recent studies, the existing literature has highlighted the fact that educational institutions are lagging behind other organizations in terms of total quality culture.
[39]	2019	This research represents a survey done at two Swedish universities, and it attempts to determine teacher educators' use of digitalization technologies and the resulting demand for digital competence in higher education. Digital competence involves, among other factors, acquiring and familiarizing with various digital tools and apps to utilize Internet and digital technology is a critical and educative approach.
[40]	2019	The digital transformation strategy seeks to create the capacity to fully use the potential of new technologies in a fast and innovative manner in the future. The study proves that planning and implementing infrastructure allow all students and staff to effectively communicate, share information, and collaborate in research skills, thus improving teaching and learning and supporting administrative functions, students, and staff to use computer systems to boost their digital skills.
[41]	2020	This work represents a bibliometric of 1590 papers from the Scopus database. In the education sector, the Digital transformation has necessitated the implementation of a long-term management strategy. The authors conclude that HEIs are progressing in managing their economic, environmental, and social sustainability, concerning digital transformation to reach the model of an open, digital, innovative, and connected institution.
[42]	2022	Higher education institutions around the world have used numerous quality management strategies. As mentioned in this work, with the growing interest of the quality measures for sustained growth in education, the potential of developing a paradigm that reflects the challenges of higher education while including comprehensive quality and social responsibility ought to be considered. Therefore, the authors proposed an approach that connects the TQM and social responsibility of organizations and higher education institutions.

Recently, knowledge management by the company has constituted a sustainable competitive advantage, forming a common point with the objectives of the quality approach: obtaining a competitive advantage [47]. Not only knowledge

management, but the implementation of business process management also helps organizations enhance their capacities through individual knowledge resources and greater collective knowledge of the organization.

The decision to manage the laboratory knowledge needs to set a standardized process by redesigning the actual processes. The following subsections define knowledge management and business process re-engineering.

### C. Knowledge Management

Knowledge is a valuable, scarce and non-substitutable resource that enables an organization to gain a sustainable competitive advantage [38]. It is a set of experiences, values, information and ideas to assess and integrate new knowledge and experiences. Knowledge is an intangible attribute that is practically impossible to simulate and is considered a strategic asset that has to be effectively managed by every organization [48]. It can be explicit or tacit.

Knowledge management is the process of getting the right knowledge to the right person at the right time. Moreover, knowledge management aims to explicit the tacit knowledge by systematizing large sets of knowledge and gathering individual knowledge [49], [50]. The purpose is to produce valuable knowledge, fulfill the knowledge demands of customers, perform knowledge and innovations and strengthen the basic competitiveness of an organization [51]. The KM approach is the integration of individuals, methods and technologies implied in planning and implementing the infrastructure of the educational institutions [52].

Therefore, the process of knowledge implies four steps: creation, retrieval/storage, transfer and application [53]. The different components of KM as cited in these works [53] in Table VII include five elements:

TABLE VII. ELEMENTS OF KM

<b>Knowledge Creation</b>	The organization's ability to create and communicate knowledge in its services, and systems. The process of knowledge creation consists in capturing a part of tacit knowledge and transforming it into explicit knowledge.
<b>Knowledge Application</b>	The most important step in knowledge management is to ensure that knowledge is productively applicable for the organization's benefit, aiming to maximize performance.
<b>Knowledge Sharing</b>	Knowledge sharing is a range of behaviors that include sharing information or helping others to inspire innovative behavior.
<b>Knowledge Capitalization (Storage)</b>	It consists in identifying its crucial knowledge, preserving it and making it sustainable while ensuring that it is shared and used by the greatest number of people. Without this capitalization effort, collective knowledge does not exist.
<b>Knowledge Transfer</b>	It is the process of transferring knowledge between individuals, groups or organizations using various means or channels of communication.

D. Business Process Re-engineering – BPR

Today, quality is identified by the process approach of the activities. The management of processes is a strategy adopted by organizations that want to become more efficient; it would make them more efficient in executing their processes and ultimately more competitive [54]. The term of ‘process approach’ first appeared in its 2000 version: "the application of the process system within an organization, as well as the identification, interactions and management of these processes". This definition summarizes the requirements of a management system of quality, and treats the customer satisfaction by adding value to each process.

The process approach has been described and established as the quality management basis in organizations through the ISO 9001 version 2015 standard. It allows to identify, map the processes, and understand their interactions in an organization [55]. In addition, according to FD X50-176 standard, “Process management can be applied to all types of organizations regardless of their size, activity, and to the various management systems implemented (quality, safety, environment, etc.)”[56].

E. Use case: Research Laboratory LAROSERI

1) *Description:* The LAROSERI research laboratory was created in 2014. It belongs to the Computer Science Department in the Faculty of Science in El Jadida, Morocco. LAROSERI includes four research teams as described in Table VIII.

2) *Challenges:* The main objective of this study is to reposition the Chouaib Doukkali University in scientific research, by applying the Total Quality Management and using a Knowledge Management System. The SWOT analysis described in Table IV has determined the factors that led us to adopt a digital transformation. The Fig. 1 shows some Laboratory challenges.

Aiming to address these issues, we propose an approach based on Total Quality Management techniques that can provide adequate solutions to digitalize manage and steer the entire laboratory. The following section presents the proposed approach.

TABLE VIII. RESEARCH LABORATORY DESCRIPTION

Department	Name of the laboratory	Research Teams
Computer sciences	Research Laboratory in Optimization, Emerging Systems, Networks and Imaging (LROSERI)	Optimization, Intelligent System and Imaging
		intelligent transportation systems
		Business Intelligence, Network and Imaging
		Decision and Information Systems

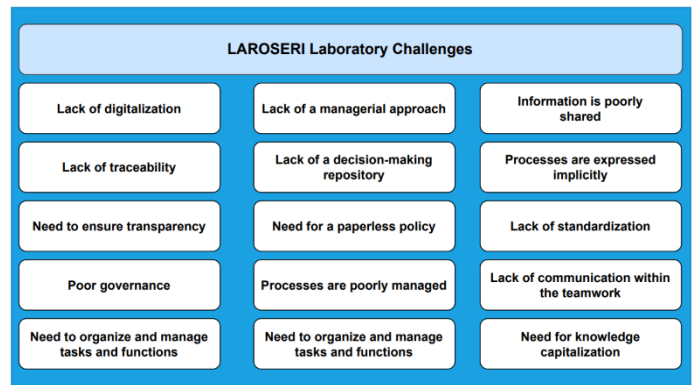


Fig. 1. Research Laboratory Challenges.

III. PROPOSED APPROACH

Related works have highlighted the potential that the TQM application in the research structure context offers. Our case study consists of proposing an innovative approach to properly digitalize and manage the research laboratory.

The proposed approach can be divided in the workflow as shown in Fig. 2. This digital transition impacts the overall functioning of companies and disrupts working methods and processes, which requires the use of managerial approaches for a long-term vision to remain competitive, efficient, and modern[57]–[59].

Fig. 2 and Fig. 3 describe the steps and tools that we used to adopt a digital transformation in our research laboratory.

A. Knowledge Management using KMS-TQM Platform

The knowledge management concerns the way knowledge is stocked and arranged. From a managerial perspective, the capitalization of knowledge is a major element in the improvement of performance via the establishing of a trustworthy resource base completed by appropriate software, and able to offer appropriate decision support. The objective of this function is to store relevant knowledge that assists actors in their operations [50].

In this work, we propose to use a KMS (Knowledge Management System) to capitalize and store knowledge in a repository, which is specifically designed for a scientific research laboratory, in order to provide relevant decision support to the different research laboratory’s actors.

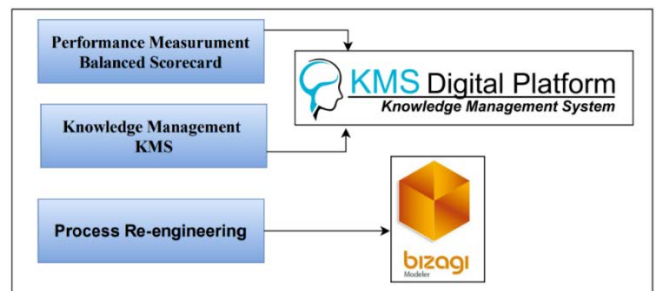


Fig. 2. Used Tools.

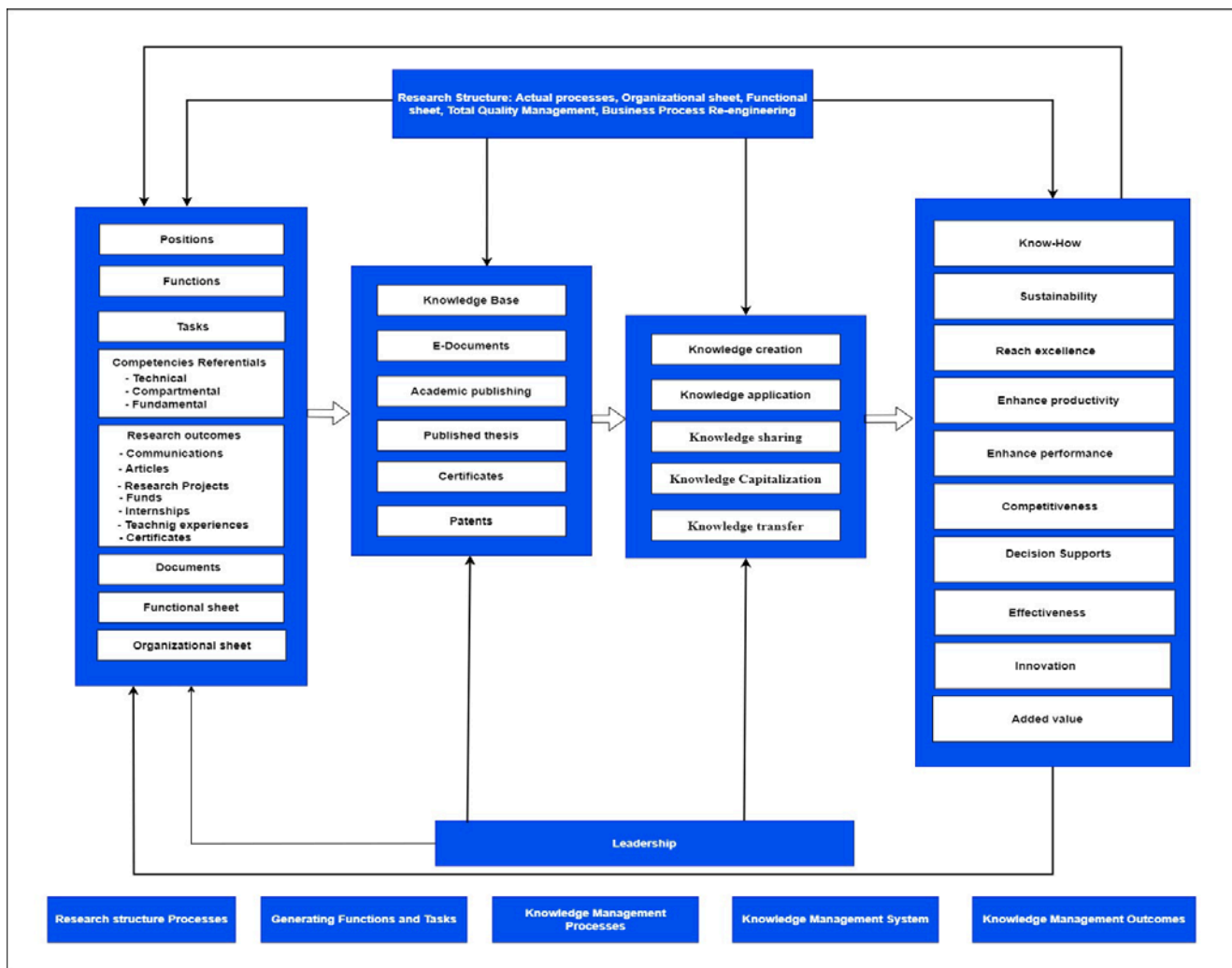


Fig. 3. Approach Workflow.

1) *KMS-TQM digital platform*: We are going to use KMS-TQM Digital platform<sup>2,3</sup> as described in the Table IX:

**B. Business Process Re-engineering using Bizagi Modeler**

Unfortunately, for Quality module, the platform offers only process mapping. Therefore, we chose the Bizagi Modeler tool for process modeling in order to make explicit all the existing processes within the laboratory. The choice of this robust process management tool aims to redesign the research laboratory process, thus integrating all actors, functions, tasks, etc.

Bizagi<sup>4</sup> is a free BPM tool (for a single user) to create, optimize and publish a process. It also provides a cloud-based collaboration environment, offering powerful and fast drag-and-drop design tools. In addition, it allows users to review process models from any location and on any device and to provide real-time feedback.

The KMS-TQM digital platform allows us the features shown in Fig. 4.

TABLE IX. KMS-TQM DIGITAL PLATFORM

<b>KMS-Digital Platform</b>	A facilitator and accelerator of digital transformation and CSR transition, thanks to its transversal approach focused on the company's business processes.
	A data collector and organizer, which prepares organizations for the next Artificial Intelligence revolution.
	A powerful and adaptable tool for steering (Balanced ScoreCard, dynamic dashboards) and monitoring (risk management, project coordination, etc.) all of the operational activities, and which has the capacity to integrate those of the stakeholders (via the integrated ISO 26000 and BCorp certification reference systems).
	With a strong digital component (digitalizing procedures, workflows, data collection), it will also support organizations in their technological transformation.

<sup>2</sup> <http://37.187.48.129:9191/062021ic-canada-certified-trainings/>

<sup>3</sup> <https://masoda.ch/>

<sup>4</sup> <https://www.bizagi.com/en/products/bpm-suite/modeler>

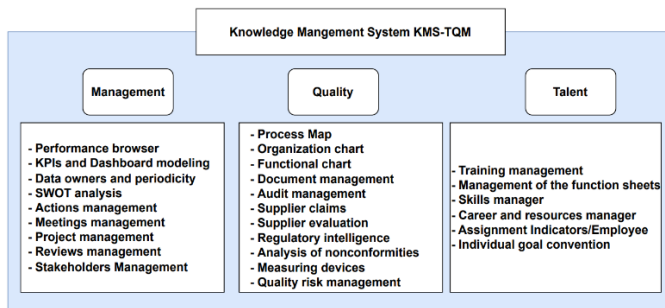


Fig. 4. Research Laboratory Digitalization and Management using KMS-TQM Digital Platform.

In this work, we intend to digitalize all the actual processes and make them as a standard in our Laboratory.

#### IV. RESULTS AND DISCUSSION

In this section, we will apply the proposed approach, first by using KMS-TQM digital platform to digitalize and manage knowledge, then by using the Bizagi Modeler to manage the laboratory processes.

The purpose of capitalizing knowledge in our case—is to make explicit functions, tasks, and competencies. It answers the question: who did what and how?

By using the KMS-TQM platform, we generate the function sheets of each position in the research laboratory. We have assigned to each position a function(s) that is linked to numerous tasks and competencies.

The first step is to create the organizational entities and assign each to a parent entity. For instance, in our case, a doctoral studies center depends on the doctoral college as mentioned in Fig. 5.

The next step is to create competencies referential to tasks. Then, assigning each task to an appropriate function becomes possible, and defining the function first manager and assignments. Fig. 6-8 below illustrate the entity's management, tasks, and function management.

After preparing functions with their owners, the platform proposes to generate a functional or organizational sheet. It summarizes all the functions depending on their relationship. Fig. 9 describes an example of the functional sheet.

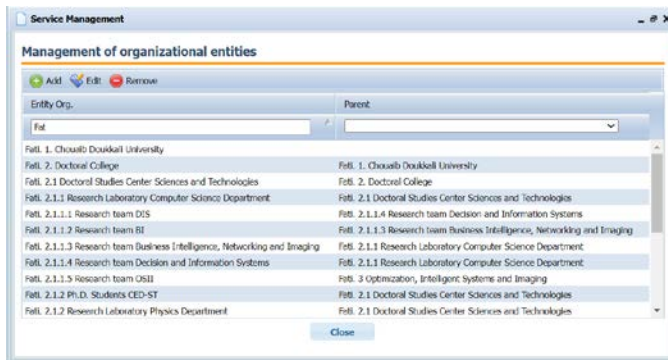


Fig. 5. Entities Management.

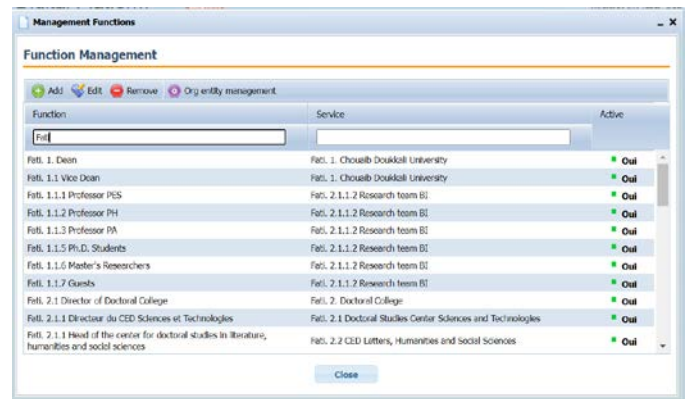


Fig. 6. Functions Management.

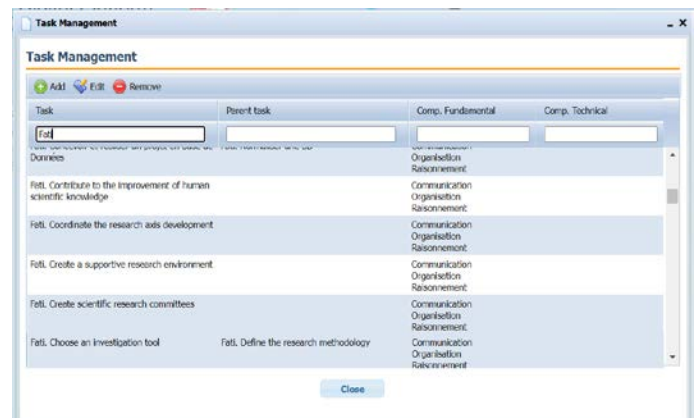


Fig. 7. Tasks Management.

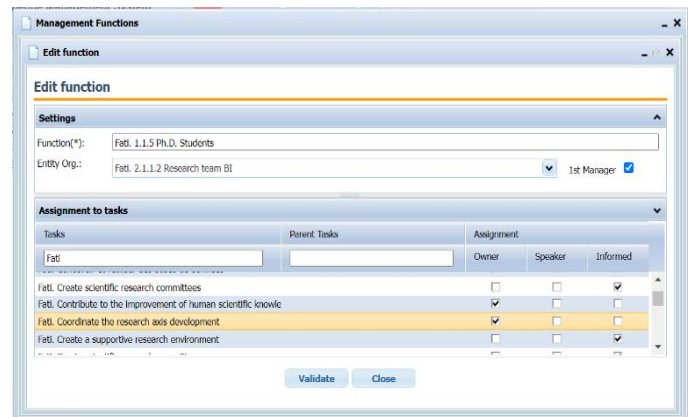


Fig. 8. Affecting Tasks to Function.

To set standardized processes within the Laboratory, each actor should elaborate a draft of the different activities; it concerns the laboratory chief, the research supervisors, research teams' chiefs, PhD students. After these tasks, the Director of the CeDoc (Center for Doctoral Studies) collects information concerning each process, capitalizes knowledge, redesigns the actual processes, and finally, validates, and standardizes these processes.

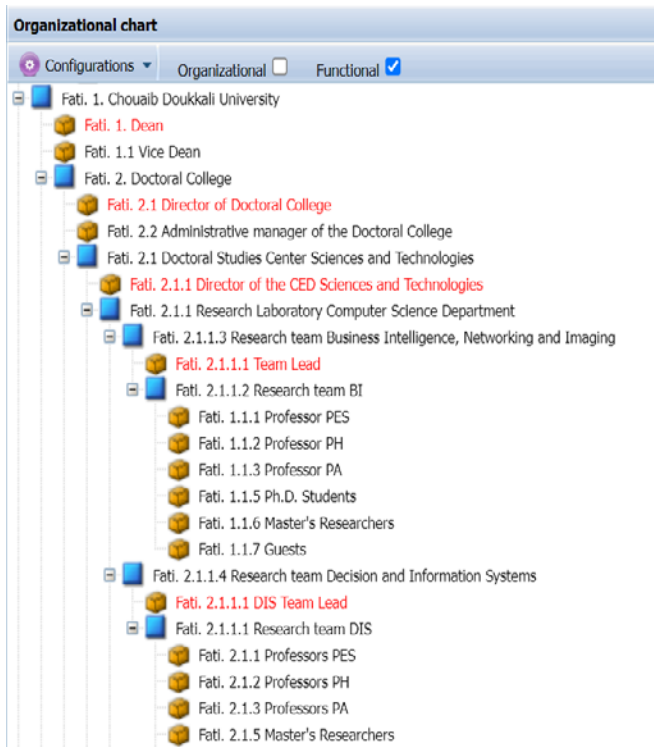


Fig. 9. An Example of a Functional Sheet.

Fig. 10 and Fig. 11 depict this activity and the different interactions.

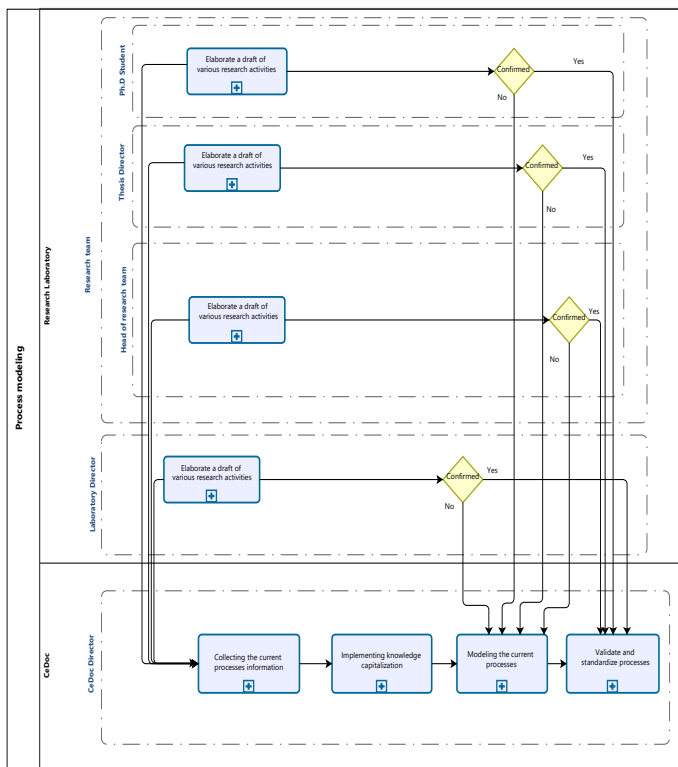


Fig. 10. Research Laboratory Processes Re-engineering.

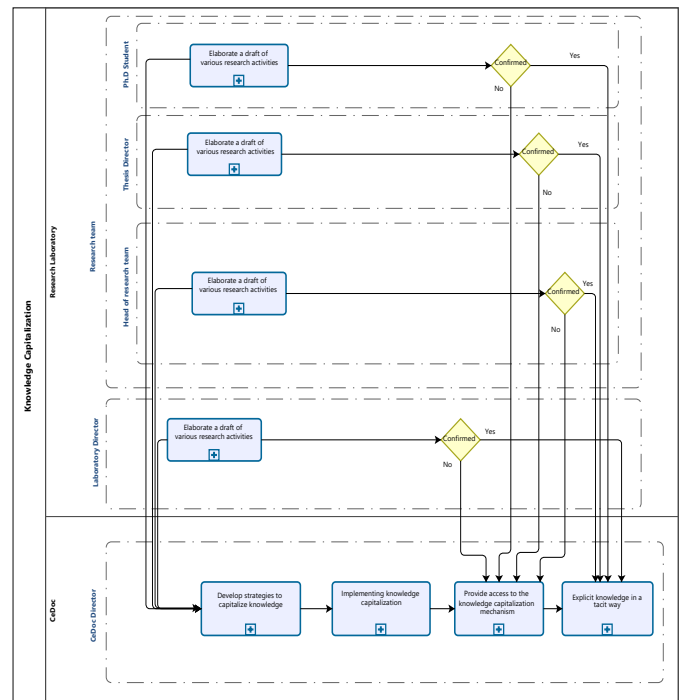


Fig. 11. Research Laboratory Knowledge Capitalization.

## V. CONCLUSION

In summary, the TQM applications aiming to digitalize and manage research structures are still very weak. This paper analyzes the factors to adopt a digital transformation in the research laboratory by applying the SWOT analysis and gives an overview of some available recent studies that apply this approach in the same context. The main objective of our research is to integrate a Knowledge Management System and apply TQM aspects, which can apply to any research laboratory that desires to adopt a digital transformation. The proposed approach briefly describes the steps to be followed to digitalize and manage knowledge and redesign the current processes.

Despite the advantages of the research approach, some limitations can be addressed through future research; the platform used in the implementation does not link process modeling with other modules, which means integration issues. Therefore, to propose a suitable and adaptable solution, it will be more practical to think of an open-source solution with huge possibilities.

In future work, as mentioned, we intend to conduct a comparative study between laboratory management information systems (LIMS) by selecting specific criteria to propose an adequate digital framework. The framework will combine the quality management system (QMS) aspects and integrate innovative solutions to digitalize and manage research structures.

## ACKNOWLEDGMENT

This work was funded by the National Center for Scientific and Technical Research (CNRST-Rabat).



REFERENCES

- [1] S. Kraus, P. Jones, N. Kailer, A. Weinmann, N. Chaparro-Banegas, et N. Roig-Tierno, « Digital Transformation: An Overview of the Current State of the Art of Research », *SAGE Open*, vol. 11, no 3, p. 21582440211047576, juill. 2021, doi: 10.1177/21582440211047576.
- [2] G. N. Zanon, A. L. Szejka, et E. de F. R. Loures, « Towards an Integrated MCDM and BSC Method to Support the Digital Transformation Strategy in Railway Companies », in *Advances in Transdisciplinary Engineering*, L. Newnes, S. Lattanzio, B. R. Moser, J. Stjepandić, et N. Wognum, Éd. IOS Press, 2021. doi: 10.3233/ATDE210109.
- [3] L. Benavides, J. Tamayo Arias, M. Arango Serna, J. Branch Bedoya, et D. Burgos, « Digital Transformation in Higher Education Institutions: A Systematic Literature Review », *Sensors*, vol. 20, no 11, p. 3291, juin 2020, doi: 10.3390/s20113291.
- [4] A. F. Teixeira, M. J. A. Gonçalves, et M. de L. M. Taylor, « How Higher Education Institutions Are Driving to Digital Transformation: A Case Study », *Educ. Sci.*, vol. 11, no 10, Art. no 10, oct. 2021, doi: 10.3390/educsci11100636.
- [5] O. El Gannour, S. Hamida, B. Cherradi, A. Raihani, et H. Moujahid, « Performance Evaluation of Transfer Learning Technique for Automatic Detection of Patients with COVID-19 on X-Ray Images », in *2020 IEEE 2nd International Conference on Electronics, Control, Optimization and Computer Science (ICECOCS)*, Kenitra, Morocco, déc. 2020, p. 1-6. doi: 10.1109/ICECOCS50124.2020.9314458.
- [6] H. Moujahid et al., « Combining CNN and Grad-Cam for COVID-19 Disease Prediction and Visual Explanation », *Intell. Autom. Soft Comput.*, vol. 32, no 2, p. 723-745, 2022, doi: 10.32604/iasc.2022.022179.
- [7] H. Moujahid, B. Cherradi, M. Al-Sarem, et L. Bahatti, « Diagnosis of COVID-19 Disease Using Convolutional Neural Network Models Based Transfer Learning », in *Innovative Systems for Intelligent Health Informatics*, vol. 72, F. Saeed, F. Mohammed, et A. Al-Nahari, Éd. Cham: Springer International Publishing, 2021, p. 148-159. doi: 10.1007/978-3-030-70713-2\_16.
- [8] S. Hamida, O. El Gannour, B. Cherradi, A. Raihani, H. Moujahid, et H. Ouajji, « A Novel COVID-19 Diagnosis Support System Using the Stacking Approach and Transfer Learning Technique on Chest X-Ray Images », *J. Healthc. Eng.*, vol. 2021, p. 1-17, nov. 2021, doi: 10.1155/2021/9437538.
- [9] O. El Gannour et al., « Concatenation of Pre-Trained Convolutional Neural Networks for Enhanced COVID-19 Screening Using Transfer Learning Technique », *Electronics*, vol. 11, no 1, p. 103, déc. 2021, doi: 10.3390/electronics11010103.
- [10] O. E. Gannour, S. Hamida, S. Saleh, Y. Lamalem, B. Cherradi, et A. Raihani, « COVID-19 Detection on X-Ray Images using a Combining Mechanism of Pre-trained CNNs », *Int. J. Adv. Comput. Sci. Appl.*, vol. 13, no 6, 2022, doi: 10.14569/IJACSA.2022.0130668.
- [11] O. El Gannour, B. Cherradi, S. Hamida, M. Jebbari, et A. Raihani, « Screening Medical Face Mask for Coronavirus Prevention using Deep Learning and AutoML », in *2022 2nd International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET)*, Meknes, Morocco, mars 2022, p. 1-7. doi: 10.1109/IRASET52964.2022.9737903.
- [12] Ø. Tønnessen, A. Dhir, et B.-T. Flåten, « Digital knowledge sharing and creative performance: Work from home during the COVID-19 pandemic », *Technol. Forecast. Soc. Change*, vol. 170, p. 120866, 2021.
- [13] S. Hamida, B. Cherradi, O. Terrada, A. Raihani, H. Ouajji, et S. Laghmati, « A Novel Feature Extraction System for Cursive Word Vocabulary Recognition using Local Features Descriptors and Gabor Filter », in *2020 3rd International Conference on Advanced Communication Technologies and Networking (CommNet)*, Marrakech, Morocco, sept. 2020, p. 1-7. doi: 10.1109/CommNet49926.2020.9199642.
- [14] S. Hamida, B. Cherradi, O. El Gannour, O. Terrada, A. Raihani, et H. Ouajji, « New Database of French Computer Science Words Handwritten Vocabulary », in *2021 International Congress of Advanced Technology and Engineering (ICOTEN)*, Taiz, Yemen, juill. 2021, p. 1-5. doi: 10.1109/ICOTEN52080.2021.9493438.
- [15] S. Hamida, B. Cherradi, et H. Ouajji, « Handwritten Arabic Words Recognition System Based on HOG and Gabor Filter Descriptors », in *2020 1st International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET)*, Meknes, Morocco, avr. 2020, p. 1-4. doi: 10.1109/IRASET48871.2020.9092067.
- [16] E. Abad-Segura, M.-D. González-Zamar, J. C. Infante-Moro, et G. Ruipérez García, « Sustainable management of digital transformation in higher education: Global research trends », *Sustainability*, vol. 12, no 5, p. 2107, 2020.
- [17] P. Krpálek, K. Berková, A. Kubišová, K. K. Krellová, D. Frendlovská, et D. Spiesová, « Formation of Professional Competences and Soft Skills of Public Administration Employees for Sustainable Professional Development », *Sustainability*, vol. 13, no 10, p. 5533, mai 2021, doi: 10.3390/su13105533.
- [18] O. R. Mahdi et I. A. Nassar, « The Business Model of Sustainable Competitive Advantage through Strategic Leadership Capabilities and Knowledge Management Processes to Overcome COVID-19 Pandemic », *Sustainability*, vol. 13, no 17, p. 9891, sept. 2021, doi: 10.3390/su13179891.
- [19] K. AKODADI, A. AAROUD, et F.-E. AIT BENNACER, « Steering the performance of academic institutions: Proposal for a KPI system for a university research laboratory in Morocco », *Altern. Manag. Econ.*, vol. 3, p. 216-237, nov. 2021, doi: 10.48374/IMIST.PRSM/AME-V3I4.28905.
- [20] M. Ingaldi et D. Klimecka-Tatar, « Digitization of the service provision process - requirements and readiness of the small and medium-sized enterprise sector », *Procedia Comput. Sci.*, vol. 200, p. 237-246, janv. 2022, doi: 10.1016/j.procs.2022.01.222.
- [21] S. Gupta, T. Tuunanen, A. K. Kar, et S. Modgil, « Managing digital knowledge for ensuring business efficiency and continuity », *J. Knowl. Manag.*, févr. 2022, doi: 10.1108/JKM-09-2021-0703.
- [22] D. Schallmo, C. A. Williams, et L. Boardman, « DIGITAL TRANSFORMATION OF BUSINESS MODELS — BEST PRACTICE, ENABLERS, AND ROADMAP », *Int. J. Innov. Manag.*, vol. 21, no 08, p. 1740014, déc. 2017, doi: 10.1142/S136391961740014X.
- [23] K. Daneshjoovash, P. Jafari, et A. Khamseh, « Commercialization cycle of entrepreneurial ideas in high-technology firms », *J. Innov. Creat. Hum. Sci.*, vol. 10, no 3, p. 41-68, févr. 2021.
- [24] S. M. Lee, D. Lee, et Y. S. Kim, « The quality management ecosystem for predictive maintenance in the Industry 4.0 era », *Int. J. Qual. Innov.*, vol. 5, no 1, p. 4, déc. 2019, doi: 10.1186/s40887-019-0029-5.
- [25] European Commission. Joint Research Centre. Institute for Prospective Technological Studies., *Digital competence in practice: an analysis of frameworks*. LU: Publications Office, 2012. Consulté le: 17 mars 2022. [En ligne]. Disponible sur: <https://data.europa.eu/doi/10.2791/82116>.
- [26] D. Schallmo, *Geschäftsmodelle erfolgreich entwickeln und implementieren: mit Aufgaben, Kontrollfragen und Templates, 2., Überarbeitete und erweiterte Auflage*. Berlin [Heidelberg]: Springer Gabler, 2018. doi: 10.1007/978-3-662-57605-2.
- [27] É. Blanc, « Une communication des organisations comme facteur de protection des risques psychosociaux liés à l'acculturation au numérique (Groupe La Poste) », *Commun. Organ. Rev. Sci. Francoph. En Commun. Organ.*, no 49, 2016.
- [28] P. Soto-Acosta, « COVID-19 pandemic: Shifting digital transformation to a high-speed gear », *Inf. Syst. Manag.*, vol. 37, no 4, p. 260-266, 2020.
- [29] V. B. Klein et J. L. Todesco, « COVID - 19 crisis and SMEs responses: The role of digital transformation », *Knowl. Process Manag.*, vol. 28, no 2, p. 117-133, 2021.
- [30] I. Avdeeva, T. Golovina, et A. Polyinin, « Change management strategy for the activities of business organizations », in *SHS Web of Conferences*, 2021, vol. 90, p. 01003.
- [31] B. Sharma et M. A. Rahim, « TQM and HRM: an integrated approach to organizational success », *J. Comp. Int. Manag.*, vol. 24, no 1, p. 27-41, 2021.
- [32] R. Hchaichi, « The Key Success Factors of Total Quality Management Implementation in State-Owned Enterprise », *Int. J. Public Adm.*, p. 1-12, 2021.



- [33] G. Bevan, « The medical therapy of peptic ulcer », *Postgrad. Med. J.*, vol. 51, no 5 Suppl, p. 14-18, 1975.
- [34] C.-K. Chen, L. Reyes, J. Dahlgaard, et S. M. Dahlgaard-Park, « From quality control to TQM, service quality and service sciences: a 30-year review of TQM literature », *Int. J. Qual. Serv. Sci.*, 2021.
- [35] Y.-S. Ho, Y. Cavacece, A. Moretta Tartaglione, et A. Douglas, « Publication performance and trends in Total Quality Management research: a bibliometric analysis », *Total Qual. Manag. Bus. Excell.*, p. 1-34, 2022.
- [36] Y. El manzani, « L'effet de la synergie entre management de la qualité et orientation marché sur l'innovation produit des entreprises marocaines certifiées ISO 9001 », These de doctorat, Lyon, 2019. Consulté le: 8 mars 2022. [En ligne]. Disponible sur: <http://www.theses.fr/2019LYSE3020>.
- [37] M. Lemtaoui Et H. Chatki, « La Relation Entre Les Pratiques Du Tqm, L'innovation Produit, L'innovation Processus Et L'innovation Organisationnelle: Revue De Litterature Et Cadre Theorique », *Rev. D'Etudes En Manag. Finance D'Organisation*, vol. 3, no 8, 2019.
- [38] M. G. Antunes, P. R. Mucharreira, M. do R. T. Justino, et J. T. Quirós, « Total quality management implementation in portuguese higher education institutions », *Multidiscip. Digit. Publ. Inst. Proc.*, vol. 2, no 21, p. 1342, 2018.
- [39] L. Amhag, L. Hellström, et M. Stigmar, « Teacher educators' use of digital tools and needs for digital competence in higher education », *J. Digit. Learn. Teach. Educ.*, vol. 35, no 4, p. 203-220, 2019.
- [40] Y. Limani, E. Hajrizi, L. Stapleton, et M. Retkoceri, « Digital Transformation Readiness in Higher Education Institutions (HEI): The Case of Kosovo », *IFAC-Pap.*, vol. 52, no 25, p. 52-57, janv. 2019, doi: 10.1016/j.ifacol.2019.12.445.
- [41] E. Abad-Segura, M.-D. González-Zamar, J. C. Infante-Moro, et G. Ruipérez García, « Sustainable Management of Digital Transformation in Higher Education: Global Research Trends », *Sustainability*, vol. 12, no 5, 2020, doi: 10.3390/su12052107.
- [42] T. Nogueiro, M. Saraiva, et F. Jorge, « Total Quality Management and Social Responsibility an Approach Through Their Synergies in Higher Education Institutions », in *Perspectives and Trends in Education and Technology*, vol. 256, A. Mesquita, A. Abreu, et J. V. Carvalho, Éd. Singapore: Springer Singapore, 2022, p. 311-321. doi: 10.1007/978-981-16-5063-5\_26.
- [43] N. V. K. Jasti, V. Venkateswaran, S. Kota, et K. S. Sangwan, « A literature review on total quality management (models, frameworks, and tools and techniques) in higher education », *TQM J.*, 2021.
- [44] Z. S. Nadim et A. H. Al-Hinai, « Critical success factors of TQM in higher education institutions context », *Int. J. Appl. Sci. Manag.*, vol. 1, no 2, p. 147-156, 2016.
- [45] L. Fatma, A. Zouari, et Abdellatif, « Proposition d'un cadre d'intégration du KM et du TQM: vers une meilleure performance de l'entreprise », présenté à LOGISTIQUA 2016: 9ème colloque international sur la logistique, mai 2016. Consulté le: 18 mars 2022. [En ligne]. Disponible sur: <https://hal.archives-ouvertes.fr/hal-01893140>
- [46] B. Yazdani, « TQM, employee outcomes and performance: the contingency effect of environmental uncertainty », *Int. J. Qual. Reliab. Manag.*, 2021.
- [47] E. G. Carayannis, J. J. M. Ferreira, et C. Fernandes, « A prospective retrospective: conceptual mapping of the intellectual structure and research trends of knowledge management over the last 25 years », *J. Knowl. Manag.*, vol. 25, no 8, p. 1977-1999, oct. 2021, doi: 10.1108/JKM-07-2020-0581.
- [48] K. K. Lim, P. K. Ahmed, et M. Zairi, « Managing for quality through knowledge management », *Total Qual. Manag.*, vol. 10, no 4-5, p. 615-621, juill. 1999, doi: 10.1080/0954412997596.
- [49] Y. Tang et C. Zhang, « Cases on Research Support Services in Academic Libraries: Peking University Library », in *Cases on Research Support Services in Academic Libraries*, IGI Global, 2021, p. 247-265.
- [50] M. Umar, M. Wilson, et J. Heyl, « The structure of knowledge management in inter-organisational exchanges for resilient supply chains », *J. Knowl. Manag.*, 2021.
- [51] L. Xiao, « Innovative application of knowledge management in organizational restructuring of academic libraries: A case study of Peking University Library », *IFLA J.*, vol. 46, no 1, p. 15-24, mars 2020, doi: 10.1177/0340035219892289.
- [52] A. Bereznoy, D. Meissner, et V. Scuotto, « The intertwining of knowledge sharing and creation in the digital platform based ecosystem. A conceptual study on the lens of the open innovation approach », *J. Knowl. Manag.*, 2021.
- [53] M. S. Mahrinasari et al., « The impact of decision-making models and knowledge management practices on performance », *Acad. Strateg. Manag. J.*, vol. 20, p. 1-13, 2021.
- [54] A.-S. Thelisson et V. Kin, « Influence des managers d'interfaces lors d'un changement organisationnel: cas d'un processus d'intégration post-fusion », *Rech. En Sci. Gest.*, no 3, p. 57-82, 2021.
- [55] « ISO 9001:2015(fr), Systèmes de management de la qualité — Exigences ». <https://www.iso.org/obp/ui/fr/#iso:std:iso:9001:ed-5:v2:fr> (consulté le 18 mars 2022).
- [56] « FD X50-176 », Afnor EDITIONS. <https://www.boutique.afnor.org/fr-fr/norme/fd-x50176/outils-de-management-management-des-processus/fa137236/25925> (consulté le 18 mars 2022).
- [57] V. Sima, I. G. Gheorghe, J. Subić, et D. Nancu, « Influences of the industry 4.0 revolution on the human capital development and consumer behavior: A systematic review », *Sustainability*, vol. 12, no 10, p. 4035, 2020.
- [58] P. F. Borowski, « Digitization, digital twins, blockchain, and industry 4.0 as elements of management process in enterprises in the energy sector », *Energies*, vol. 14, no 7, p. 1885, 2021.
- [59] S. Verma et A. Gustafsson, « Investigating the emerging COVID-19 research trends in the field of business and management: A bibliometric analysis approach », *J. Bus. Res.*, vol. 118, p. 253-261, sept. 2020, doi: 10.1016/j.jbusres.2020.06.057.

# Analyzing the Relationship between the Personality Traits and Drug Consumption (Month-based user Definition) using Rough Sets Theory

Manasik M. Nour<sup>1\*</sup>, H. A. Mohamed<sup>2</sup>  
Department of Mathematics  
College of Science and Humanities  
Prince Sattam bin Abdulaziz University  
Al-Kharj 11942, Saudi Arabia

Sumayyah I. Alshber<sup>3</sup>  
Department of Mathematics  
College of Education in Al-Dilam  
Prince Sattam bin Abdulaziz University  
Al-Kharj 11942, Saudi Arabia

**Abstract**—There is no doubt that the use of drugs has significant consequences for society, it introduces risk into the human life and causing earlier mortality and morbidity. Being a conscientious member of society, we must go ahead to prevent these young minds from life-threatening addiction. Owing to the computational complexity of wrapper approaches, the poor performance of filtering techniques, and the classifier dependency of embedded approaches, artificial intelligence and machine learning systems can provide useful tools for raising the prediction rate of drug users. Recently, the psychologists approved the recent personality traits Five Factor Model (FFM) for understanding human individual differences. The aim of this work is to propose a rough sets theory based method to investigate the relationship between drug user/non-user (month-based user definition) and the personality traits. The data of five factor personality profiles, impulsivity, sensation-seeking and biographical information of users of 21 different types of legal and illegal drugs are used to fetch all reducts and finally a set of classification rules are created to predict the drug user/non-user(month-based user definition). The outcomes demonstrate the novelty of the current work which can be summarized as The set of generalized classification rules which pronounced with logic functions build a knowledge base with excellent accuracy to analyze drug misuse successfully and may be worthy in many applications.

**Keywords**—Classification; personality traits; five factor model; rules extraction; drug abuse detection; rough sets theory; feature selection

## I. INTRODUCTION

One of the extreme serious matters taking into account the mental health in these days is drug addiction, where it has the ability to devastate life and a nation readily for their toxic and addictive effects. Drug intemperance means" the picking of diverse drugs illegally and being addicted to those drugs". Drug intemperance has turned into a severe truth for which the young descents from all lifestyles are influenced silently. Dissatisfaction is the cause for this intemperance, unemployed matters, political outburst, non-attendance of homely relationships, and non-attendance of ardent love fellowship which offers rise to disappointments [1]. Drug is having been thought to be one of the extreme used psychoactive substances. as stated by world health organization, drug consuming leads to the death of three million as well as 5.1% of several universal diseases all over the world yearly [2]. The practical

importance of the issue of estimating individual's risk of intemperance drugs is very high [3]. The connection of personality traits to risk of intemperance drugs is a continuous problem [4]. Many studies had been done to find the answer of the following Questions -"How do personality, gender, education, nationality, age, and other attributes affect this risk? Is this dependence different for different drugs? Which personality traits are the most important for evaluation of the risk of consumption of a particular drug, and are these traits different for different drugs? Is the prediction of drugs usage by a person helpful to prevent the persons from getting addicted to drugs?" Also, some related works had been done by researchers on drugs and addiction predictions to improve the methods which are used. Bergh [5] proposed a way to Predicting Alcohol Consumption in Adolescents from Historical Text Messaging Data. Belcher et al. [6] studied the personality traits and sensitivity or resilience to drug intemperance. Weissman, et al. [7] studied the effects of the drug intemperance adolescent and it is found that there is a strong connection between reward and cognitive control brain networks. Andreassen et al. [8] studied the relevance between behavioral addictions and the FFM of personality. Kumar, et. al. [9] proposed efficient prediction of drug-drug interaction using deep learning models. In this work all the questions which posed above have been reformulated as classification problem and an effective data mining technique dependent on rough set theory was employed to address these issues and extracting classification rules to predict the Drug User/Non-User (month-based user definition).

## II. RESEARCH PROBLEM

Psychologists tried many times to identify the connections between personality traits and drug user/non-user. Many studies are done and data mining techniques and methodologies had been used to manage these issues such as decision trees, linear discriminant analysis, and statistics estimation techniques [10]. The main aim of this work is to find answers to these questions: Which personality traits have the great importance for estimation of the risk of abusing drugs, and are these traits different for different drugs? What are the effects of personality, gender, education, nationality, age, and other factors on abusing drugs? What about this dependence for several drugs?

\* Corresponding Author

### A. Personality Traits

In recent years and due to the development in scientific research, the psychologists approved the recent personality traits Five Factor Model (FFM) for realization of human individual variances [11]. It consists of Neuroticism (N), Extraversion (E), Openness to Experience (O), Agreeableness (A), and Conscientiousness (C). These traits can be defined as follow:

- N : "Neuroticism is a long-term tendency to experience negative emotions such as nervousness, tension, anxiety and depression (associated adjectives [10]: anxious, self-pitying, tense, touchy, unstable, and worrying)"
- E: " Extraversion manifested in characters who are outgoing, warm, active, assertive, talkative, and cheerful; these persons are often in search of stimulation (associated adjectives: active, assertive, energetic, enthusiastic, outgoing, and talkative)".
- O: "Openness to experience is associated with a general appreciation for art, unusual ideas, and imaginative, creative, unconventional, and wide interests (associated adjectives: artistic, curious, imaginative, insightful, original, and wide interest)".
- A: "Agreeableness is a dimension of interpersonal relations, characterized by altruism, trust, modesty, kindness, compassion and cooperativeness (associated adjectives: appreciative, forgiving, generous, kind, sympathetic, and trusting)".
- C: "Conscientiousness is a tendency to be organized and dependable, strong-willed, persistent, reliable, and efficient (associated adjectives: efficient, organised, reliable, responsible, and thorough)".

The values of the five factors (N, E, O, A, C) are utilized as inputs in various statistical methodologies for prediction, diagnosis, and risk estimation. These methodologies and techniques are used a wide range of fields where personality has a great importance such as medicine, psychology, psychiatry, education, sociology, and many others areas. Other two additional feature of personality confirmed to be leading for analysis of matter use, Impulsivity (Imp) and Sensation-Seeking (SS) [12].

- Imp: "Impulsivity is defined as a tendency to act without adequate forethought "
- SS: "Sensation-Seeking is defined by the search for experiences and feelings, that are varied, novel, complex and intense, and by the readiness to take risks for the sake of such experiences"

### B. Rough Sets Theory

Rough sets theories (RST) is the core of most recent approximations based mathematical model to investigate the imprecision and uncertainty present in knowledge [13-17], as well as extract decision rules which act as classification scheme for prediction. We can say that it is a tool for data mining or knowledge discovery in relational databases. It is a formal approximation of a crisp set defined by its two

approximations namely, Upper and Lower approximation [18] as shown in Fig. 1.

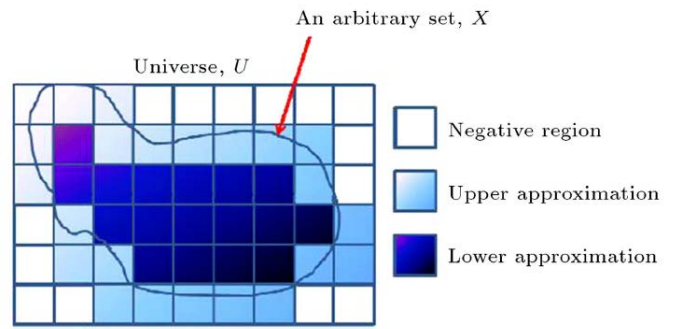


Fig. 1. Represent Tation of a Set Approximation of an Arbitrarily Set X in U.

The definition of the indiscernible relation IND(B) is:

$$IND(B) = \{ (x, y) \in U \mid \forall a \in B, a(x) = a(y) \} \quad (1)$$

Also, in decision system  $(U, A)$  let  $B \subseteq A$  and  $X \subseteq U$ , the lower approximate  $\underline{B}(x)$ , upper approximate  $\overline{B}(x)$  and the boundary of X denoted by BND(X) are written as:

$$\underline{B}(x) = \{x \in U \mid [x]_B \subseteq X\} \quad (2)$$

$$\overline{B}(x) = \{x \in U \mid [x]_B \cap X \neq \emptyset\} \quad (3)$$

$$BND(X) = \overline{B}(x) - \underline{B}(x) \quad (4)$$

The B-positive region of X and The B-negative region of X, denoted as  $POS_B(X)$ ,  $NEG_B(X)$  respectively are written as:

$$POS_B(X) = \underline{B}X \quad (5)$$

$$NEG_B(X) = U - \underline{B}X \quad (6)$$

The accuracy of approximation can be written as:

$$\alpha_B(X) = \frac{|BX|}{|BX|} \quad (7)$$

Where  $|x|$  is the cardinality of X. Obviously  $0 \leq \alpha_B(X) \leq 1$ . The rough membership function can be written as

$$\mu_X^B(x) = \frac{|X \cap [x_i]_{Ind(B)}|}{|[x_i]_{Ind(B)}|} \quad (8)$$

Obviously,

$$\mu_x^B(x) \in [0,1] \tag{9}$$

### III. ANALYSIS

In the life of the any human there are various factors (attributes) for addiction that lead to increase the probability of drug consumption. Some of these attributes correlated with psychological, social, environmental, and economic characteristics [19, 20]. The most important risk factors are likewise associated with personality traits [21]. So this study proposes a methodology based on rough set theory to extract decision rules for predicting drug user/non-user (month-based user definition). We defined different categories (classifications) of “drug users” based on the regency of use as follows: class of “non-users”, “year-based”, “month-based” and “week-based” user/non-user.

Linear discriminates for user/non-user separation is evaluated by several methods, here we will consider the following Relations:

For users:

$$Th + \sum k_i CT > 0 \tag{10}$$

For non users:

$$Th + \sum k_i CT \leq 0 \tag{11}$$

Where

*Th* : is the thresholds.

*CT* : is the conditional attributes.

*k<sub>i</sub>* : are the coefficients of the conditional attributes.

Data had been taken from the database which was collected by Elaine Fehrman [22] for 21 different types of legal and illegal drugs separately, where the values of the five factors (N, E, O, A, C) in addition to Impulsivity (Imp) and Sensation-Seeking (SS) as well as biographical data: age, gender, and education are used as the conditional attributes in the decision table shown in Table I.

Now, we will use rough sets methodology to find structural connections within the given data to obtain all reducts and

finally a set of generalized classification rules are extracted to predict the drug user/non-user. The overall steps of the suggested rough sets methodology are shown in Fig. 2.

By using RST analysis toolkit software called ROSETTA where Semi-Naïve algorithm were used to discretize the data in Table I to be as shown in Table II where “ \* means do not care condition” . After that reduction techniques based rough sets is used to determine the minimal reducts of (factors) attributes that can characterize all the knowledge in the decision tables as presented in Table III. Finally, the knowledge gained from all extracted reducts can be outlined by rough sets dependency rules as shown in Table IV.

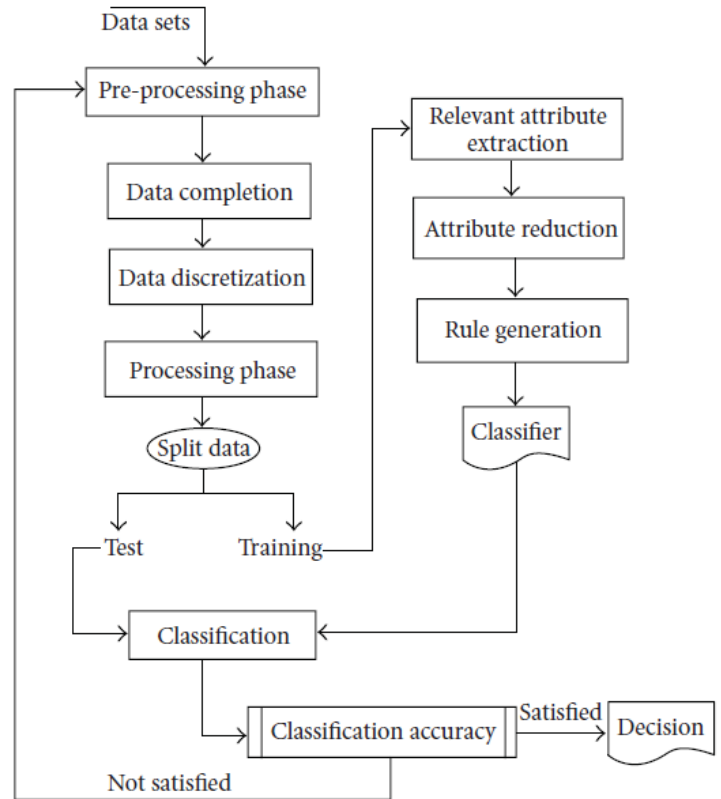


Fig. 2. The Overall Steps of the Suggested Rough Sets Methodology.

TABLE I. DECISION TABLE OF COEFFICIENTS OF LINEAR DISCRIMINANT FOR USER/NON-USER (YEAR-BASED USER DEFINITION)

	TH	Age	Gndr	Edu	N	E	O	A	C	Imp	SS	Drug
x1	0.130	0.263	0.058	0.590	0.096	0.588	0.111	0.078	0.083	0.193	0.402	Alcohol
x2	0.543	0.643	0.293	0.249	0.063	0.176	0.347	0.103	0.201	0.241	0.418	Amphetamines
x3	0.821	0.361	0.365	0.229	0.223	0.114	0.178	0.144	0.018	0.088	0.749	Amyl nitrite
x4	0.416	0.115	0.292	0.243	0.711	0.128	0.284	0.180	0.072	0.167	0.418	Benz.
x5	0.122	0.542	0.250	0.394	0.132	0.166	0.547	0.037	0.132	0.015	0.355	Cannabis
x6	0.132	0.138	0.501	0.284	0.161	0.107	0.379	0.004	0.440	0.488	0.193	Chocolate
x7	0.597	0.624	0.270	0.029	0.345	0.212	0.007	0.305	0.054	0.062	0.523	Cocaine
x8	0.273	0.019	0.042	0.369	0.261	0.637	0.043	0.035	0.239	0.424	0.385	Caffeine

x9	0.836	0.154	0.555	0.131	0.449	0.114	0.075	0.253	0.156	0.076	0.581	Crack
x10	0.633	0.820	0.257	0.047	0.139	0.093	0.284	0.123	0.165	0.028	0.328	Ecstasy
x11	1.037	0.560	0.226	0.371	0.181	0.350	0.159	0.397	0.016	0.368	0.154	Heroin
x12	0.793	0.776	0.386	0.020	0.097	0.147	0.340	0.098	0.268	0.039	0.139	Ketamine
x13	0.693	0.519	0.467	0.224	0.012	0.190	0.409	0.136	0.240	0.022	0.427	Legal highs
x14	0.851	0.722	0.284	0.173	0.006	0.045	0.541	0.007	0.032	0.098	0.252	LSD
x15	0.551	0.404	0.296	0.259	0.262	0.443	0.417	0.270	0.105	0.003	0.399	Methadone
x16	0.764	0.594	0.267	0.233	0.184	0.236	0.604	0.019	0.066	0.070	0.239	MMushrooms
x17	0.019	0.461	0.283	0.530	0.092	0.037	0.318	0.013	0.375	0.157	0.389	Nicotine
x18	1.027	0.785	0.222	0.003	0.081	0.059	0.174	0.215	0.073	0.096	0.482	VSA
x19	0.545	0.525	0.364	0.187	0.301	0.180	0.224	0.295	0.013	0.129	0.530	Heroin pleiad
x20	0.019	0.576	0.241	0.339	0.133	0.207	0.514	0.098	0.172	0.073	0.355	Ecstasy pleiad
x21	0.346	0.309	0.380	0.254	0.479	0.125	0.274	0.215	0.123	0.171	0.534	Benz. pleiad

TABLE II. DECISION TABLE OF COEFFICIENTS OF LINEAR DISCRIMINANT FOR USER/NON-USER (YEAR-BASED USER DEFINITION)

	TH	Age	Gndr	Edu	N	E	O	A	C	Imp	SS	Drug
x1	[0.075, 0.131)	[-0.286, -0.189)	[0.050, 0.280)	[0.480, *)	[0.094, 0.139)	[0.400, 0.613)	[-0.144, -0.093)	[0.057, *)	[-0.094, -0.074)	[-0.340, -0.145)	[0.401, 0.410)	Alcohol
x2	[-0.544, -0.479)	[-0.682, -0.633)	[-0.294, -0.292)	[-0.251, -0.246)	[0.035, 0.072)	[-0.178, -0.171)	[0.344, 0.363)	[-0.113, -0.100)	[-0.220, -0.186)	[0.206, 0.305)	[0.410, 0.423)	Amphetamines
x3	[-0.828, -0.807)	[-0.382, -0.335)	[-0.372, -0.364)	[-0.231, -0.226)	[-0.231, -0.203)	[-0.119, -0.079)	[-0.119, -0.079)	[-0.162, -0.140)	[-0.025, -0.002)	[-0.093, -0.063)	[0.665, *)	Amyl nitrite
x4	[-0.479, -0.381)	[-0.189, -0.067)	[-0.292, -0.288)	[-0.246, -0.238)	[-0.595, *)	[-0.137, -0.126)	[0.279, 0.301)	[-0.197, -0.162)	[0.063, 0.073)	[0.162, 0.169)	[0.410, 0.423)	Benz.
x5	[-0.234, -0.070)	[-0.551, -0.533)	[-0.253, -0.245)	[-0.462, -0.382)	[-0.132, -0.114)	[-0.171, -0.156)	[0.544, 0.576)	[-0.067, -0.028)	[-0.148, -0.127)	[0.009, 0.019)	[0.342, 0.370)	Cannabis
x6	[0.131, 0.203)	[0.060, 0.146)	[0.280, *)	[-0.311, -0.271)	[-0.172, -0.150)	[0.100, 0.160)	[0.363, 0.394)	[-0.011, 0.002)	[-0.011, 0.002)	[-0.011, 0.002)	[0.174, 0.216)	Chocolate
x7	[-0.615, -0.574)	[-0.633, -0.609)	[-0.276, -0.268)	[0.025, 0.038)	[0.323, 0.397)	[0.160, 0.400)	[-0.025, 0.076)	[-0.351, -0.300)	[0.035, 0.063)	[0.042, 0.066)	[0.503, 0.527)	Cocaine
x8	[0.203, *)	[-0.067, 0.060)	[0.090, 0.050)	[0.208, 0.480)	[0.221, 0.262)	[0.613, *)	[-0.059, -0.025)	[0.024, 0.057)	[-0.239, -0.220)	[0.396, *)	[0.370, 0.387)	Caffeine
x9	[-0.843, -0.828)	[0.146, *)	[-0.511, *)	[-0.152, -0.064)	[0.397, 0.464)	[-0.119, -0.079)	[-0.093, -0.059)	[-0.261, -0.234)	[0.115, *)	[0.075, 0.086)	[0.558, 0.665)	Crack
x10	[-0.663, -0.615)	[-0.568, -0.551)	[-0.262, -0.253)	[0.038, 0.208)	[-0.150, -0.136)	[0.076, 0.100)	[0.279, 0.301)	[-0.129, -0.113)	[-0.168, -0.148)	[-0.033, -0.012)	[0.290, 0.342)	Ecstasy
x11	[-0.807, -0.778)	[-0.780, -0.749)	[-0.426, -0.383)	[0.012, 0.025)	[-0.114, -0.054)	[-0.156, -0.137)	[0.076, 0.167)	[-0.100, -0.067)	[-0.321, -0.254)	[-0.063, -0.033)	[0.147, 0.174)	Heroin
x12	[-0.807, -0.778)	[-0.780, -0.749)	[-0.426, -0.383)	[0.012, 0.025)	[-0.114, -0.054)	[-0.156, -0.137)	[0.076, 0.167)	[-0.100, -0.067)	[-0.321, -0.254)	[-0.063, -0.033)	[0.147, 0.174)	Ketamine
x13	[-0.728, -0.663)	[-0.522, -0.490)	[-0.511, -0.426)	[-0.226, -0.205)	[-0.054, -0.003)	[-0.198, -0.185)	[0.394, 0.413)	[-0.140, -0.129)	[-0.254, -0.239)	[0.019, 0.042)	[0.423, 0.455)	Legal highs
x14	[-0.939, -0.843)	[-0.749, -0.682)	[-0.288, -0.283)	[-0.180, -0.152)	[-0.003, 0.035)	[-0.079, -0.004)	[0.528, 0.544)	[0.002, 0.010)	[-0.049, -0.025)	[-0.145, -0.093)	[0.246, 0.290)	LSD
x15	[-0.574, -0.548)	[-0.432, -0.382)	[-0.330, -0.294)	[-0.271, -0.256)	[0.262, 0.282)	[-0.396, 0.396)	[0.413, 0.466)	[-0.282, -0.261)	[-0.114, -0.094)	[-0.012, 0.009)	[0.394, 0.401)	Methadone
x16	[-0.778, -0.728)	[-0.609, -0.585)	[-0.268, -0.262)	[-0.238, -0.231)	[-0.203, -0.172)	[-0.293, -0.221)	[0.576, *)	[-0.028, -0.011)	[-0.074, -0.049)	[0.066, 0.072)	[0.216, 0.246)	MMushrooms
x17	[-0.070, 0.000)	[-0.490, -0.432)	[-0.283, -0.276)	[-0.064, 0.462)	[0.087, 0.094)	[-0.004, 0.048)	[0.301, 0.329)	[0.010, 0.024)	[-0.407, -0.321)	[0.143, 0.162)	[0.387, 0.394)	Nicotine
x18	[-1.032, -0.939)	[-0.802, -0.780)	[-0.224, -0.090)	[-0.064, 0.012)	[0.072, 0.087)	[0.048, 0.076)	[0.167, 0.199)	[-0.234, -0.197)	[0.073, 0.115)	[0.086, 0.113)	[0.455, 0.503)	VSA
x19	[-0.548, -0.544)	[-0.533, -0.522)	[-0.364, -0.330)	[-0.205, -0.180)	[0.282, 0.323)	[-0.185, -0.178)	[0.199, 0.249)	[-0.300, -0.282)	[-0.002, 0.015)	[0.113, 0.143)	[0.527, 0.532)	Heroin pleiad
x20	[0.000, 0.075)	[-0.585, -0.568)	[-0.245, -0.233)	[-0.355, -0.311)	[-0.136, -0.132)	[-0.221, -0.198)	[0.466, 0.528)	[-0.100, -0.067)	[-0.186, -0.168)	[0.072, 0.075)	[0.342, 0.370)	Ecstasy pleiad
x21	[-0.381, -0.234)	[-0.335, -0.286)	[-0.383, -0.372)	[-0.256, -0.251)	[0.464, 0.595)	[-0.126, -0.119)	[0.249, 0.279)	[-0.234, -0.197)	[-0.127, -0.114)	[0.169, 0.206)	[0.532, 0.558)	Benz. pleiad

TABLE III. REDUCTS OF DISCRETIZED DECISION TABLE

<b>Reduct</b>	{ Imp }	{ TH }	{ C }	{ Age }	{ Gndr }	{ Edu }	{ N }
<b>Support</b>	100	100	100	100	100	100	100
<b>Length</b>	1	1	1	1	1	1	1
<b>Reduct</b>	{ E, O }	{ E, SS }	{ A, SS }	{ O,A }	{ E,A }	{ O, SS }	
<b>Support</b>	100	100	100	100	100	100	
<b>Length</b>	2	2	2	2	2	2	

TABLE IV. THE SET OF GENERATED RULES

<b>Rule</b>	<b>LHS Support</b>	<b>RHS Support</b>	<b>RHS Accuracy</b>	<b>LHS Coverage</b>	<b>RHS Stability</b>
O([-0.025, 0.076]) AND SS([0.503, 0.527]) => Drug (Cocaine)	1	1	1.0	0.043478	1.0
O([-0.059, -0.025]) AND SS([0.370, 0.387]) => Drug (Caffeine)	1	1	1.0	0.043478	1.0
O([-0.093, -0.059]) AND SS([0.558, 0.665]) => Drug (Crack)	1	1	1.0	0.043478	1.0
O([0.279, 0.301]) AND SS([0.290, 0.342]) => Drug (Ecstasy)	1	1	1.0	0.043478	1.0
O([0.076, 0.167]) AND SS([0.147, 0.174]) => Drug (Heroin)	1	1	1.0	0.043478	1.0
O([0.329, 0.344]) AND SS([*, 0.147]) => Drug (Ketamine)	1	1	1.0	0.043478	1.0
O([0.301, 0.329]) AND SS([0.387, 0.394]) => Drug (Nicotine)	1	1	1.0	0.043478	1.0
O([0.167, 0.199]) AND SS([0.455, 0.503]) => Drug (VSA)	1	1	1.0	0.043478	1.0
O([0.199, 0.249]) AND SS([0.527, 0.532]) => Drug (Heroin pleiad)	1	1	1.0	0.043478	1.0
E([-0.178, -0.171]) AND A([-0.113, -0.100]) => Drug (Amphetamines)	1	1	1.0	0.043478	1.0
E([-0.119, -0.079]) AND A([-0.162, -0.140]) => Drug (Amyl nitrite)	1	1	1.0	0.043478	1.0
E([-0.137, -0.126]) AND A([-0.197, -0.162]) => Drug (Benz.)	1	1	1.0	0.043478	1.0
E([-0.171, -0.156]) AND A([-0.067, -0.028]) => Drug (Cannabis)	1	1	1.0	0.043478	1.0
E([0.100, 0.160]) AND A([-0.011, 0.002]) => Drug (Chocolate)	1	1	1.0	0.043478	1.0
E([0.160, 0.400]) AND A([-0.351, -0.300]) => Drug (Cocaine)	1	1	1.0	0.043478	1.0
E([0.613, *]) AND A([0.024, 0.057]) => Drug (Caffeine)	1	1	1.0	0.043478	1.0
O([0.076, 0.167]) AND A([*, -0.351]) => Drug (Heroin)	1	1	1.0	0.043478	1.0
O([0.329, 0.344]) AND A([-0.100, -0.067]) => Drug (Ketamine)	1	1	1.0	0.043478	1.0
O([0.394, 0.413]) AND A([-0.140, -0.129]) => Drug (Legal highs)	1	1	1.0	0.043478	1.0
O([0.528, 0.544]) AND A([0.002, 0.010]) => Drug (LSD)	1	1	1.0	0.043478	1.0
O([0.413, 0.466]) AND A([-0.282, -0.261]) => Drug (Methadone)	1	1	1.0	0.043478	1.0
O([0.576, *]) AND A([-0.028, -0.011]) => Drug (MMushrooms)	1	1	1.0	0.043478	1.0
O([0.301, 0.329]) AND A([0.010, 0.024]) => Drug (Nicotine)	1	1	1.0	0.043478	1.0
O([0.167, 0.199]) AND A([-0.234, -0.197]) => Drug (VSA)	1	1	1.0	0.043478	1.0
O([0.199, 0.249]) AND A([-0.300, -0.282]) => Drug (Heroin pleiad)	1	1	1.0	0.043478	1.0
O([0.466, 0.528]) AND A([-0.100, -0.067]) => Drug (Ecstasy pleiad)	1	1	1.0	0.043478	1.0
O([0.249, 0.279]) AND A([-0.234, -0.197]) => Drug (Benz. pleiad)	1	1	1.0	0.043478	1.0
A([0.057, *]) AND SS([0.401, 0.410]) => Drug (Alcohol)	1	1	1.0	0.043478	1.0
A([-0.113, -0.100]) AND SS([0.410, 0.423]) => Drug (Amphetamines)	1	1	1.0	0.043478	1.0
A([-0.162, -0.140]) AND SS([0.665, *]) => Drug (Amyl nitrite)	1	1	1.0	0.043478	1.0
A([-0.197, -0.162]) AND SS([0.410, 0.423]) => Drug (Benz.)	1	1	1.0	0.043478	1.0
A([-0.351, -0.300]) AND SS([0.503, 0.527]) => Drug (Cocaine)	1	1	1.0	0.043478	1.0
A([0.024, 0.057]) AND SS([0.370, 0.387]) => Drug (Caffeine)	1	1	1.0	0.043478	1.0



A([-0.261, -0.234]) AND SS([0.558, 0.665]) => Drug (Crack)	1	1	1.0	0.043478	1.0
A([-0.129, -0.113]) AND SS([0.290, 0.342]) => Drug (Ecstasy)	1	1	1.0	0.043478	1.0
A([*, -0.351]) AND SS([0.147, 0.174]) => Drug (Heroin)	1	1	1.0	0.043478	1.0
A([-0.100, -0.067]) AND SS([*, 0.147]) => Drug (Ketamine)	1	1	1.0	0.043478	1.0
A([-0.140, -0.129]) AND SS([0.423, 0.455]) => Drug (Legal highs)	1	1	1.0	0.043478	1.0
E ([-0.396, -0.293]) AND SS([0.147, 0.174]) => Drug (Heroin)	1	1	1.0	0.043478	1.0
E ([-0.156, -0.137]) AND SS([*, 0.147]) => Drug (Ketamine)	1	1	1.0	0.043478	1.0
E ([-0.198, -0.185]) AND SS([0.423, 0.455]) => Drug (Legal highs)	1	1	1.0	0.043478	1.0
E ([-0.079, -0.004]) AND SS([0.246, 0.290]) => Drug (LSD)	1	1	1.0	0.043478	1.0
E ([-0.004, 0.048]) AND SS([0.387, 0.394]) => Drug (Nicotine)	1	1	1.0	0.043478	1.0
E ([-0.156, -0.137]) AND O([0.329, 0.344]) => Drug (Ketamine)	1	1	1.0	0.043478	1.0
E ([-0.198, -0.185]) AND O([0.394, 0.413]) => Drug (Legal highs)	1	1	1.0	0.043478	1.0
E ([-0.079, -0.004]) AND O([0.528, 0.544]) => Drug (LSD)	1	1	1.0	0.043478	1.0
E ([*, -0.396]) AND O([0.413, 0.466]) => Drug (Methadone)	1	1	1.0	0.043478	1.0
E ([-0.293, -0.221]) AND O([0.576, *]) => Drug (MMushrooms)	1	1	1.0	0.043478	1.0
E ([-0.004, 0.048]) AND O([0.301, 0.329]) => Drug (Nicotine)	1	1	1.0	0.043478	1.0
E ([0.400, 0.613]) AND O([-0.144, -0.093]) => Drug (Alcohol)	1	1	1.0	0.043478	1.0
E ([-0.178, -0.171]) AND O([0.344, 0.363]) => Drug (Amphetamines)	1	1	1.0	0.043478	1.0
E ([-0.119, -0.079]) AND O([*, -0.144]) => Drug (Amyl nitrite)	1	1	1.0	0.043478	1.0
N([0.094, 0.139]) => Drug (Alcohol)	1	1	1.0	0.043478	1.0
N([0.035, 0.072]) => Drug (Amphetamines)	1	1	1.0	0.043478	1.0
N([*, -0.203]) => Drug (Amyl nitrite)	1	1	1.0	0.043478	1.0
N([0.595, *]) => Drug (Benz.)	1	1	1.0	0.043478	1.0
N([-0.132, -0.114]) => Drug (Cannabis)	1	1	1.0	0.043478	1.0
N([-0.172, -0.150]) => Drug (Chocolate)	1	1	1.0	0.043478	1.0
N([0.323, 0.397]) => Drug (Cocaine)	1	1	1.0	0.043478	1.0
N([0.221, 0.262]) => Drug (Caffeine)	1	1	1.0	0.043478	1.0
Edu([0.480, *]) => Drug (Alcohol)	1	1	1.0	0.043478	1.0
Edu([-0.251, -0.246]) => Drug (Amphetamines)	1	1	1.0	0.043478	1.0
Edu([-0.231, -0.226]) => Drug (Amyl nitrite)	1	1	1.0	0.043478	1.0
Edu([-0.246, -0.238]) => Drug (Benz.)	1	1	1.0	0.043478	1.0
Edu([0.208, 0.480]) => Drug (Caffeine)	1	1	1.0	0.043478	1.0
Edu([-0.152, -0.064]) => Drug (Crack)	1	1	1.0	0.043478	1.0
Age([-0.522, -0.490]) => Drug (Legal highs)	1	1	1.0	0.043478	1.0
Age([-0.749, -0.682]) => Drug (LSD)	1	1	1.0	0.043478	1.0
Age([-0.432, -0.382]) => Drug (Methadone)	1	1	1.0	0.043478	1.0
Age([-0.585, -0.568]) => Drug (Ecstasy pleiad)	1	1	1.0	0.043478	1.0
Age([-0.335, -0.286]) => Drug (Benz. pleiad)	1	1	1.0	0.043478	1.0
C ([-0.220, -0.186]) => Drug (Amphetamines)	1	1	1.0	0.043478	1.0
C ([-0.025, -0.002]) => Drug (Amyl nitrite)	1	1	1.0	0.043478	1.0
C ([0.063, 0.073]) => Drug (Benz.)	1	1	1.0	0.043478	1.0
C ([-0.148, -0.127]) => Drug (Cannabis)	1	1	1.0	0.043478	1.0
C ([*, -0.407]) => Drug (Chocolate)	1	1	1.0	0.043478	1.0
C ([0.035, 0.063]) => Drug (Cocaine)	1	1	1.0	0.043478	1.0
C ([0.015, 0.035]) => Drug (Heroin)	1	1	1.0	0.043478	1.0

C ((-0.321, -0.254)) => Drug (Ketamine)	1	1	1.0	0.043478	1.0
Imp((0.113, 0.143)) => Drug (Heroin pleiad)	1	1	1.0	0.043478	1.0
Imp((0.072, 0.075)) => Drug (Ecstasy pleiad)	1	1	1.0	0.043478	1.0
Imp((0.169, 0.206)) => Drug (Benz. pleiad)	1	1	1.0	0.043478	1.0
Imp(Undefined) => Drug (Undefined)	1	1	1.0	0.043478	1.0
TH((0.075, 0.131)) => Drug (Alcohol)	1	1	1.0	0.043478	1.0
TH((-0.544, -0.479)) => Drug (Amphetamines)	1	1	1.0	0.043478	1.0
TH((-0.828, -0.807)) => Drug (Amyl nitrite)	1	1	1.0	0.043478	1.0
TH((-0.615, -0.574)) => Drug (Cocaine)	1	1	1.0	0.043478	1.0
TH((0.203, *)) => Drug (Caffeine)	1	1	1.0	0.043478	1.0
TH((-0.843, -0.828)) => Drug (Crack)	1	1	1.0	0.043478	1.0
TH((-0.663, -0.615)) => Drug (Ecstasy)	1	1	1.0	0.043478	1.0
TH(*, -1.032)) => Drug (Heroin)	1	1	1.0	0.043478	1.0

TABLE V. DRUG GROUPS ACCORDING TO THE VALUES WHICH DIFFER FROM THE SAMPLE MEAN FOR GROUPS OF USERS FOR THE MONTH-BASED USER/NON-USER

Group No.		N	E	O	A	C
1	Alcohol, Chocolate, Caffeine	Neutral value (i.e. all factors for these legal drug consumers does not significantly differ from the sample mean)				
2	Nicotine	high	Neutral	high	Neutral	low
3	Amphetamines, Ketamine, and Legal highs	high	Neutral	high	low	low
4	Ecstasy and LSD	Neutral	high	high	low	low
5	Amyl nitrite	Neutral	Neutral	Neutral	low	low
6	Cannabis and Magic Mushrooms	Neutral	Neutral	high	low	low
7	Benzodiazepines, Heroin, and Methadone	high	low	high	low	low
8	Crack	high	low	Neutral	low	low
9	Cocaine and VSA	high	high	high	low	low

As shown in Table IV, the extracted decision rules represent the influence of the personality traits on the risk of drug consumption. For drug users, it is found that the N and O values of are moderately high or neutral, while the value of A and C are moderately low or neutral. In general we can call that, the risk of drug consumption increases as the values of “N” and “O” increase, while the risk decreases as there is an increase in the values of “A” and “C”. So we can conclude that drug users (month-based user definition) have higher values of on N and O, and lower on A and C when compared to drug non-users (month-based user definition). The impact of the values of “E” is cannot be generalized i.e. specific. Also, all drugs can be separated into nine groups according to the values which differ from the sample mean for groups of users for the month-based user/non-user as shown in Table V.

#### IV. CONCLUSION

This work used the principles of rough set theory to find and explain the relationship between drug use and personality traits, impulsivity, and sensation seeking, by generating a set of decision rules to investigate and predict the impact of the personality traits on drug user/Non-user (month-based user definition). It is concluded that for drug users, the N and O values of are moderately high or neutral, while the value of A

and C are moderately low or neutral. These results demonstrate the novelty of the current work which can be summarized as the suggested methodology has simplified logic-based rules required to effectively analyse drug abuse, construct a knowledge base with high accuracy to analyze drug misuse successfully and may be valuable in many applications. The future work will be extended by using other intelligent systems like neural networks, genetic algorithms, fuzzy approaches, and so forth.

#### ACKNOWLEDGMENT

“This work was supported by the Deanship of Scientific Research at Prince Sattam bin Abdulaziz University under the research project (PSAU-2022/02/20334)”

#### CONFLICTS OF INTEREST

“The authors declare no conflict of interest “

#### REFERENCES

- [1] Arif, Md Ariful Islam, Saiful Islam Sany, Farah Sharmin, Md Sadekur Rahman, and Md Tarek Habib. "Prediction of addiction to drugs and alcohol using machine learning: A case study on Bangladeshi population.", International Journal of Electrical and Computer Engineering 11, no. 5, 2021, pp. 4471-4480.

- [2] Kumari, D., and A. Swetapadma. "Analysis of alcohol abuse using improved artificial intelligence methods.", *Journal of Physics: Conference Series*, vol. 1950, no. 1, 2021, pp. 012003.
- [3] Merz, Fabien. "United Nations Office on Drugs and Crime: World Drug Report 2017. 2017.", *SIRIUS-Zeitschrift für Strategische Analysen* 2, no. 1, 2018, pp 85-86.
- [4] Kotov, Roman, Wakiza Gamez, Frank Schmidt, and David Watson. "Linking "big" personality traits to anxiety, depressive, and substance use disorders: a meta-analysis.", *Psychological bulletin* 136, no. 5, 2010, pp.768.
- [5] Bergh, Adrienne Melissa Martin. "A Machine Learning Approach to Predicting Alcohol Consumption in Adolescents From Historical Text Messaging Data.", PhD diss., Chapman University, 2019.
- [6] Belcher, Annabelle M., Nora D. Volkow, F. Gerard Moeller, and Sergi Ferré. "Personality traits and vulnerability or resilience to substance use disorders.", *Trends in cognitive sciences* 18, no. 4, 2014, pp. 211-217.
- [7] Weissman, David G., Roberta A. Schriber, Catherine Fassbender, Olivia Atherton, Cynthia Krafft, Richard W. Robins, Paul D. Hastings, and Amanda E. Guyer. "Earlier adolescent substance use onset predicts stronger connectivity between reward and cognitive control brain networks.", *Developmental cognitive neuroscience* 16, 2015, pp.121-129.
- [8] Andreassen, Cecilie Schou, Mark D. Griffiths, Siri Renate Gjertsen, Elfrid Krossbakken, Siri Kvam, and Ståle Pallesen. "The relationships between behavioral addictions and the five-factor model of personality." *Journal of behavioral addictions* 2, no. 2, 2013, pp. 90-99.
- [9] Kumar Shukla, Prashant, et al. "Efficient prediction of drug-drug interaction using deep learning models.", *IET Systems Biology* 14.4, 2020, pp. 211-216.
- [10] Fehrman, E., V. Egan, A. N. Gorban, J. Levesley, E. M. Mirkes, and A. K. Muhammad. "Personality Traits and Drug Consumption. A Story Told by Data. Cham.", 2001.
- [11] McCrae, Robert R., and Oliver P. John. "An introduction to the five factor model and its applications.", *Journal of personality* 60, no. 2, 1992, pp.175-215.
- [12] Kopstein, Andrea N., Rosa M. Crum, David D. Celentano, and Steven S. Martin. "Sensation seeking needs among 8th and 11th graders: characteristics associated with cigarette and marijuana use.", *Drug and alcohol dependence* 62, no. 3, 2001, pp. 195-203.
- [13] Z. Pawlak, " On learning—a rough set approach", In *Symposium on Computation Theory*, Springer, Berlin, Heidelberg, 1984, pp. 197-227.
- [14] H. A. Nabwey, "A Hybrid Approach for Extracting Classification Rules Based on Rough Set Methodology and Fuzzy Inference System and Its Application in Groundwater Quality Assessment.", *Advances in Fuzzy Logic and Technology*, Springer, Cham, 2017, pp. 611-625.
- [15] H. A. Nabwey, M. Modather, and M. Abdou, "Rough set theory based method for building knowledge for the rate of heat transfer on free convection over a vertical flat plate embedded in a porous medium.", *International Conference on Computing, Communication and Security (ICCCS)*, IEEE, 2015, pp. 1-8.
- [16] H. A. Nabwey, " An approach based on Rough Sets Theory and Grey System for Implementation of Rule-Based Control for Sustainability of Rotary Clinker Kiln", *International Journal of Engineering Research and Technology*, Vol. 12, No. 12, 2019, pp. 2604-2610.
- [17] H.A. Nabwey, "A method for fault prediction of air brake system in vehicles", *International Journal of Engineering Research and Technology*, Vol. 13, No. 5, 2020, pp. 1002-1008.
- [18] Arabani, M., and M. Pirouz. "Liquefaction prediction using rough set theory." *Scientia Iranica* 26, no. 2, 2019, pp. 779-788.
- [19] World Health Organization. "Prevention of mental disorders: Effective interventions and policy options: Summary report.", 2004.
- [20] Ventura, Carla AA, Jacqueline de Souza, Miyeko Hayashida, and Paulo Sérgio Ferreira. "Risk factors for involvement with illegal drugs: opinion of family members or significant others." *Journal of Substance Use* 20, no. 2, 2015, pp. 136-142.
- [21] Dubey, Charu, Meenakshi Arora, Sanjay Gupta, and Bipin Kumar. "Five Factor correlates: A comparison of substance abusers and non-substance abusers.", *Journal of the Indian Academy of Applied Psychology*, 2010.
- [22] Fehrman E, Egan V. Drug consumption, collected online March 2011 to March 2012, English speaking countries. ICPSR36536-v1. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2016-09-09. Deposited by Mirkes E. <http://doi.org/10.3886/ICPSR36536.v1>.
- [23] Fehrman, Elaine, Awaz K. Muhammad, Evgeny M. Mirkes, Vincent Egan, and Alexander N. Gorban. "The five factor model of personality and evaluation of drug consumption risk." *Data science*, 2017, pp. 231-242.

# Wavelet Multi Resolution Analysis based Data Hiding with Scanned Secrete Images

Kohei Arai

Faculty of Science and Engineering  
Saga University, Saga City, Japan

**Abstract**—Wavelet Multi Resolution Analysis (MRA) based data hiding with scanned secrete images is proposed for improvement of invisibility of the secrete images. Daubechies (biorthogonal basis was adopted as the wavelet, but it was demonstrated that the key image (or secret image data) information can be restored with the biorthogonal wavelet. Also, the information of what to adopt as the biorthogonal wavelet is hidden. Key image information can also be protected by doing so, that the horizontal biorthogonal wavelet of the image does not have to be the same as the vertical biorthogonal wavelet, and the insertion position of the secret image data can be freely selected. It is also possible to divide the bit string of the secret image data and insert it into an arbitrary high frequency component, that the information hiding capability changes depending on the number of bit strings (information amount) of the secret image data, and the secret image in the public image data. Random scanning is effective for improving the visibility of data, selection of scanning method type, random number initial value. It was shown that sharing only among parties is useful for improving confidentiality, resistance to noise, resistance to data compression, and resistance to tampering with data.

**Keywords**—Multi-Dimensional wavelet transformation; multi resolution analysis: MRA; image data hiding; scanned secrete image; Daubechies basis function; invisibility

## I. INTRODUCTION

Although personal works are often represented by digital format files, etc., the current situation is that the method of claiming the copyright of the digital contents works is unbearable. In other words, copyright cannot be protected even if it is plagiarized without knowing how to claim the copyright. The importance of digital forensics is being emphasized. That is, evidence is getting more important. How should the proof of copyright infringement be left behind? That is the question of the research. For this purpose, the digital content itself is hidden and hidden only between the recipient and the third party.

There is a method to send and receive so that it does not exist. Data hiding technology. Data hiding is a general term for steganography and digital watermarking. When the information to be embedded is important and its existence is not known, steganography, and when the content itself in which the secret information is embedded is important, it is generally referred to as a digital watermark [1].

In steganography, there is a trade-off between the quality of multimedia content and the amount of information that can be embedded. In digital watermarking, there is a trade-off

between resistance to attacks and the amount of information that can be embedded [2]. To efficiently perform a cryptographic protocol, such as a digital fingerprint system, a method that suppresses the amount of calculation and communication may be an excellent method. In addition, it is important to have a digital watermark technology that is resistant to attacks such as falsification and deletion of embedded digital fingerprints [3].

The data hiding introduced in this paper allows digital contents to keep secret keys of authors in circulation so that copyright can be claimed. This makes it possible for an author who can know the secret key to claim the copyright by taking out the distribution content from the distribution content.

The secret key must not be visible to the distributed contents, and this invisibility is important. It is also important to improve the confidentiality by devising a method to keep the secret key in the distribution contents. One of the methods is to hide the secret key in the decomposition factor in wavelet multiresolution analysis. Especially, if it is hidden in high wavelet frequency components, the visibility is generally high [4], [5], [6], [7]. The wavelet-based data hiding method includes reversible data hiding by the histogram gap method based on the integer wavelet, in addition to the method based on this multiresolution analysis [8].

Wavelets allows time-frequency analysis. Wavelet Multi Resolution Analysis: MRA based on biorthogonal basis function of Daubechies is applicable for a variety of application fields [9], [10], [11]. One of the application fields is data hiding.

If the frequency component in which the secret key is embedded is searched by the brute force method or the like, the secret key may be stolen or tampered with. Therefore, it is extremely dangerous to simply perform data hiding using multi-resolution analysis. Therefore, in this paper, data hiding by multi-resolution analysis is preprocessed, and the parameters of the preprocessing that only authors who can do it also need to know together with the information about the frequency component to be embedded. The author devised it so that the author could not find the key. To improve the invisibility of the secret key image in the distribution image, the secret key image is rescanned in accordance with the Hilbert scan algorithm or random scanning algorithm as a preprocessing of the MRA-based data hiding.

Section II outlines data hiding based on multi-resolution analysis, and Section III proposes a method for performing

sequence order conversion and permutation conversion processing by random scanning on the bit array of the secret key. The data hiding process in which the image data in the database is selected as the original image is exemplified, and the confidentiality and the visibility difficulty of the secret content in the distribution content are evaluated in Section 4. Sections 5 and 6 gives conclusions and future work, respectively.

## II. OUTLINE OF DATA HIDING BASED ON MULTI-RESOLUTION ANALYSIS

Method for data hiding based on Legall 5/2 (Cohen-Daubechies-Feauveau: CDF 5/3) wavelet with data compression and random scanning of secret imagery data is proposed [12]. Improvement of secret image invisibility in circulation image with Dyadic wavelet-based data hiding with run-length coding is also proposed [13]. Meanwhile, noble method for data hiding using Steganography Discrete Wavelet Transformation: DWT and Cryptography Triple Data Encryption Standard: DES is proposed and well reported [14].

In this paper, MRA based data hiding method with random scanning of the insert secrete image is proposed.

### A. Wavelet Multi-Resolution Analysis

The wavelet transforms of a given discrete scalar signal  $f = (f_1, f_2, \dots, f_n)^T$  is described as  $C_n f$  by a square matrix  $C_n$  composed of a sequence  $\{p_k\}$  and a sequence  $\{q_k\}$ .  $p_i$  is for low-frequency components, coefficient  $q_i$  is for high-frequency components,  $C_n$  divides  $f$  into low-frequency components and high-frequency components, and is composed of sequences  $\{a_k\}$  and sequences  $\{b_k\}$  The square matrix  $H_n$  is expressed as follows,

$$H_n C_n = I_n \quad (1)$$

where  $I_n$  is an identity matrix. And then, the following equation is defined.

$$H_n = C_n^T \quad (2)$$

When Eq. (2), the biorthogonal wavelet transform is an orthogonal wavelet transform, that is, the orthogonal wavelet transform is a kind of biorthogonal wavelet transform.

### B. 2D(Two Dimensional) Discrete Wavelet Transformation

For 2D image signals, this process is performed horizontally and vertically one level at a time. Fig. 1 shows the band components when two-dimensional DWT is performed twice. In the figure, L indicates a low frequency component, and H indicates a high frequency component. The image is decomposed into four bands (LL, LH, HL, HH) by the first two-dimensional DWT, and the lowest band component (LL) is further divided into four bands (LLLL, LLLH, LLHL, LLHH).

Following are the related research works: Data hiding method replacing LSB of hidden portion for secrete image with Run-Length coded image is proposed [15]. Meanwhile, Data hiding method with Principal Component Analysis: PCA and image coordinate conversion is proposed for improvement of invisibility of the secret key image [16].

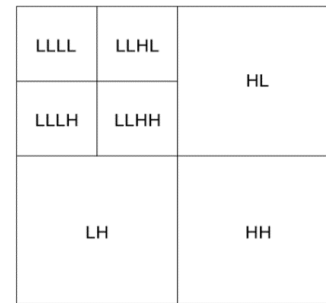


Fig. 1. Band Components after the 2D DWT.

## III. PROPOSED METHOD

When the DWT is applied to  $n$  time series data in one stage, it can be decomposed into  $n / 2$  high frequency components and  $n / 2$  low frequency components. By further subjecting the  $n / 2$  low frequency components to a one-stage DWT, the  $n / 4$  low frequency components and the  $n / 4$  high frequency components can be decomposed. By repeating this, the number of data becomes 1 or 2. This is shown in Fig. 2. This is called the Laplacian Pyramid.

In this case, the size of the image in each stage is halved both vertically and horizontally by the DWT. A Dyadic Wavelet that does not downsize is also proposed. Also, in this case, a certain low-frequency component image is decomposed into four, but a Multi Wavelet that decomposes this into 16 images is also proposed. It is reported that they are effective for noise removal and data compression, respectively. These Wavelets and many others are published in the Special Issue on Visualization Information Society of Reference [9], so please refer to them.

The original time series data can be completely restored by applying the inverse wavelet transform (Inverse DWT: IDWT) for the number of transform stages using the high frequency components and the low frequency components of each stage generated by this decomposition. Of the decomposed frequency component data, the fact that "the human eye has a low resolution of high frequency components" is used to embed the secret data into one of the high frequency components and reconstruct it with the secret data embedded. When attempting to restore to the original image level, the secret data is embedded in the high frequency component, so that data like the original image is reconstructed in a state where it is difficult to see.

The data generated in this way is called distribution data (content). The distribution contents are contents that are open to the public and can be obtained by anyone. Therefore, they are exposed to the risk of plagiarism. This distributed content is almost the same as the original content but differs from the original content in that the secret data is embedded in the high frequency component. Even if the distributed content is stolen, the copyright holder can claim the copyright by extracting and showing the secret content (e.g., copyright) embedded in the high frequency component. Fig. 3 shows a series of processing flow from embedding secret data for asserting copyright in such copyrighted content, generating a distribution image, and restoring the secret data and the original content from the distribution image.

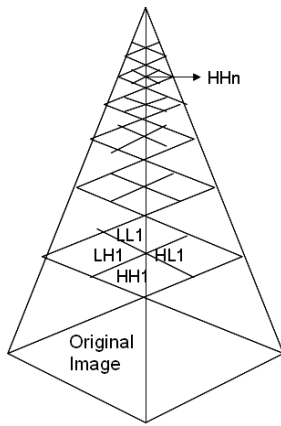


Fig. 2. Laplacian Pyramid.

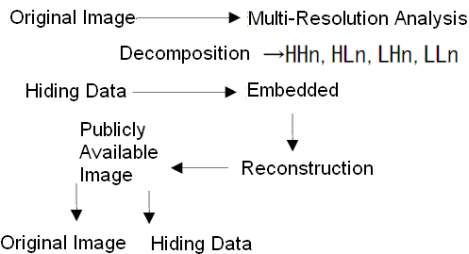


Fig. 3. Process Flow of Data Hiding based on Multi Resolution Analysis (MRA).

The visibility of the secret content in the distributed content is greatly influenced by the confidential data to be embedded (frequency component of the content) and the place to be embedded (frequency component). Therefore, as shown in Fig. 4, the location where the secret content is embedded is a very important factor in considering the visibility of the secret image content in the distribution image content. In this case, LH1 of MRA (Fig. 4 (c)) of the original image (Fig. 4 (a)) is replaced to the secrete image of “CRAMPS” (Fig. 4 (b)). Then the reconstructed image is derived by Inverse DWT.



Fig. 4. LH1 of MRA is replaced to the Hiding Image Content of “CRAMPS”.

The scanning method of the secret image data can be changed from the normal line sequential scanning to the random scanning to improve the visibility of the secret image in the distribution image. The author proposes a method to obtain a distribution image in random scan (rand) by using the support length (dbn) of Daubechies basis function used in MRA and the initial value (rand50 / 5000) of uniform random numbers used in random scan as parameters.

#### IV. EXPERIMENT

##### A. Preliminary Results

A method has also been proposed to improve the visibility of the secret data in the distribution image by scanning it again before embedding the secret data. It is premised that the rules are shared. In contrast to normal image data that is line-sequential scanning, a secret image is converted to random scanning that determines the scanning order by, for example, generating a random number. It converts and stores 2D spatial data into a dictionary array (1D data).

The conversion of this scanning method can be performed by a permutation conversion matrix. At this time, if the random number generation rule information is shared between the sending and receiving parties, the inverse matrix of the permutation conversion matrix is applied after extracting secret data in random scanning. By doing so, the reverse conversion of the scanning method becomes possible and the secret data in the line sequential scanning can be reproduced. Since it is difficult for a third party to obtain the information of the scanning method when embedding the secret data, it is difficult to obtain the secret data. The confidentiality of information is also improved.

The experimental results are as follows: The used data is the original image shown in Fig. 5 (a) (The band 3 red area in which the Thematic Mapper: TM sensor mounted on the Landsat satellite observed near the Yamato interchange of Nagasaki Highway near Saga city) The original image is composed of 128x128 pixels, the secret data is composed of 64x64 pixels, and the quantized bits are 8 in each case. Landsat / TM is a 30m spatial resolution multi-spectral scanner with spectral bands of five bands from blue to near infrared and one band in thermal infrared.

The wavelet division was applied to the original image by one stage, and the secret data was embedded in HH1. At that time, the secret data was embedded in HH1 with line sequential scanning and the random number generation method of Merthenne Twister was used. Compared with the method of generating uniform random numbers, scanning again based on that, and embedding in HH1, the distribution image obtained by reconstructing using the embedded image is almost the same as the original image.

When the author tries to reconstruct the HH1 using this method, the secret data can be restored as shown in Fig. 7 (a), (b) and (c) for line sequential scanning (raster scan), random scanning and Hilbert scan, respectively. An example of Hilbert scan is shown,



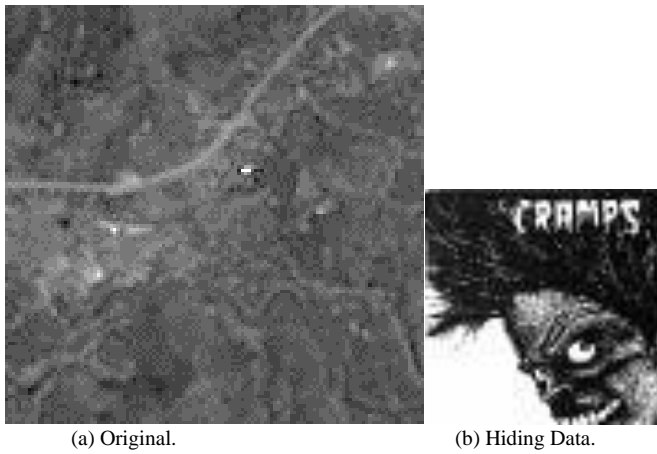


Fig. 5. Original Image of Landsat-5/TM Band 3 Data of Saga City and Hiding Data.

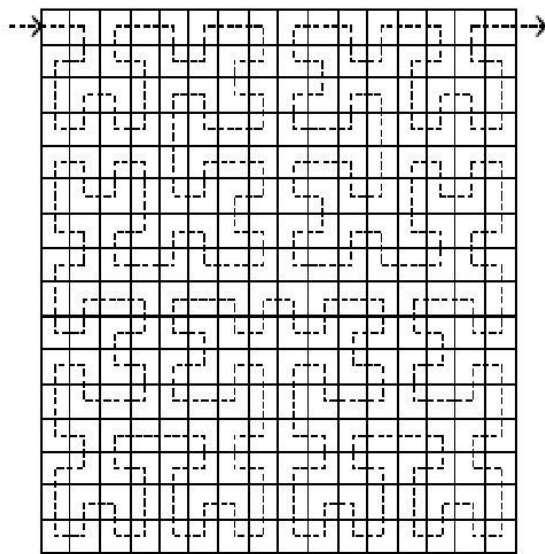


Fig. 6. Hilbert Scan.

The Hilbert curve is one of the simplest curves which pass through all points in a space (Fig. 6). Many researchers have worked on this curve from the engineering point of view, such as for an expression of two-dimensional patterns, for data compression in an image or in color space, for pseudo color image displays, etc.

If the secret data is embedded in the original image as it is, the secret data itself can be restored in HH1, but in the case of random scanning, the secret data cannot be restored without knowing the control parameters for random number generation.

On the other hand, Fig. 8 (a) shows the secret image of “CRAMPS” derived from the Hilbert scanning. Also, Fig. 8 (b) and (c) are the randomly scanned secret image and the raster scanned secret image, respectively. Not only random scan, but also Hilbert scan can be used for improvement of invisibility of the hidden secret image from the reconstructed image. Fig. 8 (d) shows the reconstructed image derived from the decomposed image embedding the secret image with Hilbert scanning. Also, Figs. 8 (e) and (f) show the reconstructed

images derived from the decomposed image embedding the secret image with raster scanning and random scanning, respectively.

These are images obtained by scanning the key image by raster, Hilbert, and random scanning, replacing the original image with HH1 after MRA, and performing wavelet transform on the reconstructed image. As is apparent from these, in the case of raster scanning, the key image itself appears and there is no confidentiality. On the other hand, in Hilbert scanning, only horizontal stripe noise appears, and in random scanning, only random noise appears, so it is difficult to visually recognize the key image.

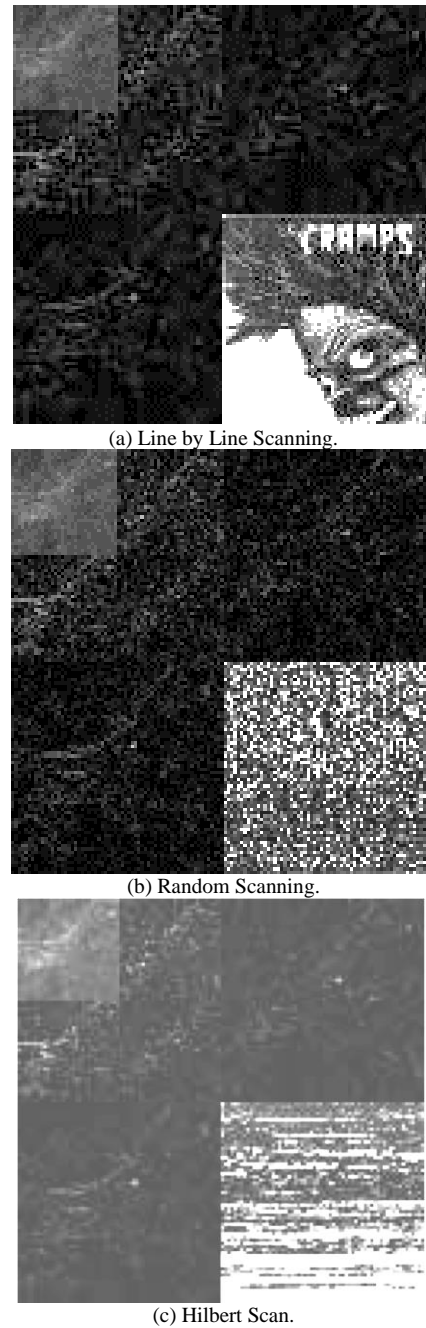
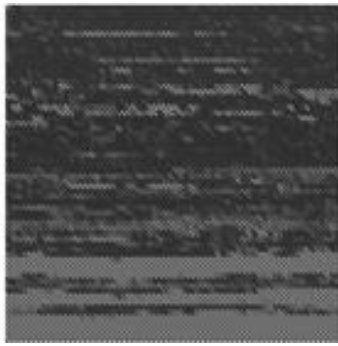
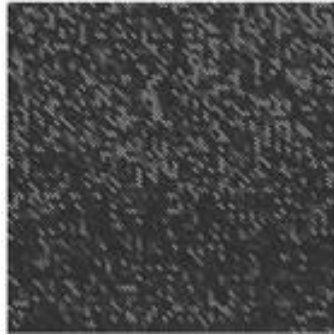


Fig. 7. Reconstructed Hiding Data from Publicly Available Image Content Derived from the MRA based Methods with the Different Scanning Schemes.



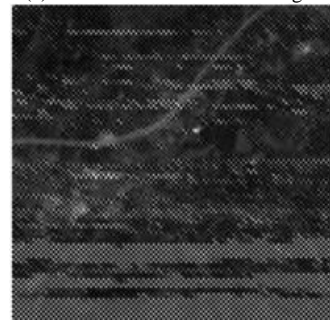
(a) Hilbert Scanned Secrete Image.



(b) Randomly Scanned Secrete Image.



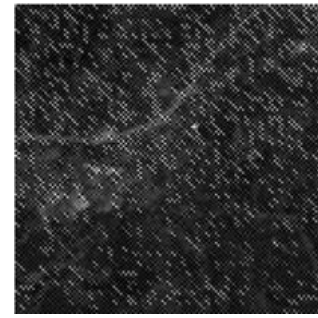
(c) Raster Scanned Secrete Image.



(b) Reconstructed Image (Hilbert Scan).



(c) Reconstructed Image Raster Scan.



(d) Reconstructed Image (Random Scan).

Fig. 8. Secrete Image of “CRAMPS” Derived from the Hilbert Scanning and the Reconstructed Image Extracted from the Decomposed Image Derived from the Decomposed Image Embedding the Secrete Image with Hilbert, Random and Raster Scanning.

Furthermore, by understanding the parameters related to the scanning order generation method in random scanning and Hilbert scanning only between the sending and receiving parties, only the parties can know the key image, and the confidentiality and confidentiality can be improved.

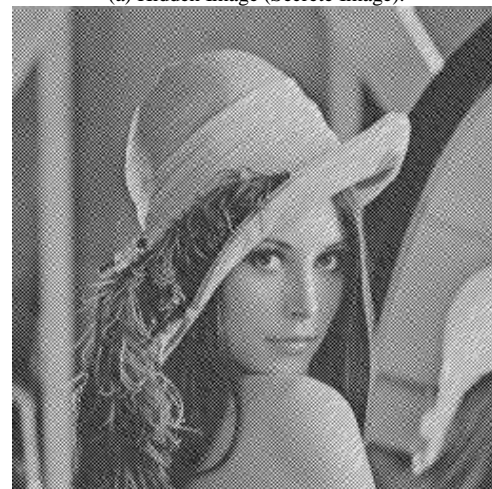
### B. Experimental Results

Fig. 9(a) is an example of a secret image. Fig. 9(b), (c), and (d) show the distribution images when this is inserted into each of HH1, HL1, and LH1 of the above-mentioned original image (Lena).

As is clear from Fig. 9, the secret image data can be visually recognized on the distribution image. As shown in Fig. 10, this secret image data is changed from line sequential scanning to random scanning to improve visibility.



(a) Hidden Image (Secrete Image).



(b) Publicly Available Reconstructed Image through Embedding the Hiding Image at HH1 Component.



(c) Publicly Available Reconstructed Image through Embedding the Hiding Image at HL1 Component.



(a) HH1.



(d) Publicly Available Reconstructed Image through Embedding the Hiding Image at LH1 Component.



(b) HL1.

Fig. 9. Hidden Image and Publicly Available Reconstructed Images through Embedding the Hiding Image at HH1, HL1 and LH1 Components.

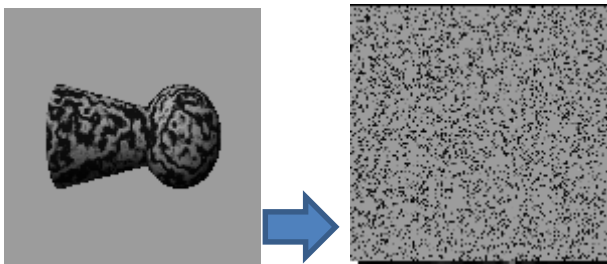


Fig. 10. Scanning Scheme Conversion from the Line-by-line to Random.

Fig. 11(a), (b), and (c) show the results of improving the visibility of the secret image data on the distribution image due to the change of the scanning method. Fig. 10(a), (b), and (c) show the circulation images when they are inserted into HH1, HL1, and LH1 of the original image (Lena).



(c) LH1.

Fig. 11. Hidden Image and Publicly Available Reconstructed Images through Embedding the Hiding Image at HH1, HL1 and LH1 Components after the Scanning Scheme Conversion for Hidden Image from Line-by-line to Random.

Distribution image and original image in line sequential scanning (Normal) and random scanning (rand) using the support length (dbn) of the Daubechies basis function used for MRA and the initial value of uniform random numbers (rand50 / 5000) used for random scanning as parameters Table I shows a comparison of the Root Mean Square Difference (RMSD) between the original and the publicly available reconstructed images and the results show that random scanning is better than line-sequential scanning, and that longer support length is better than short support length. It can be seen that the initial value of the random number is not so affected.

Since the visibility of the secret image data on the distribution image does not depend on the initial value of the random number used, if this initial value is hidden by steganography between the sending and receiving parties, only the party who knows this initial value will have the secret value. Image data can then be restored.

TABLE I. COMPARISONS OF ROOT MEAN SQUARE DIFFERENCE (RMSD) BETWEEN THE ORIGINAL AND THE PUBLICLY AVAILABLE RECONSTRUCTED IMAGES THROUGH DATA HIDING BASED ON MRA WITH EMBEDDING THE HIDING IMAGE TO HL1, HH1 AND LH1 AND WITH SCANNING SCHEME CONVERSION FROM LINE-BY-LINE TO RANDOM

Scanning Method	HL1	HH1	LH1
Normal(db2)	69.594	69.137	69.183
Normal(db4)	69.397	69.089	69.058
Normal(db8)	69.518	69.069	69.056
rand50(db2)	68.790	68.297	68.340
rand50(db4)	68.609	68.215	68.247
rand50(db8)	68.568	68.135	68.123
rand5000(db2)	68.856	68.357	68.427
rand5000(db4)	68.665	68.291	68.316
rand5000(db8)	68.633	68.182	68.202

## V. CONCLUSION

The author has introduced a method that improves the confidentiality by applying principal component transformation and oblique coordinate transformation as preprocessing for data hiding based on wavelet multiresolution analysis. The author investigated the confidentiality when a third party attempts to extract secret data from only the data for distribution.

The method introduced in this paper allows only the author who knows the characteristics of the original multispectral image to recover the secret data, i.e., when the information of the original image needs to be protected. The author also showed how to convert the scanning method of the secret data from line-sequential to random scanning, which leads to the improvement of the confidentiality of the secret data and the visibility difficulty in the distribution image. By sharing the equation parameters only between the sending and receiving parties, more confidential data hiding can be realized.

In this paper, the Daubechies basis function is adopted as the wavelet, but the secret data can be restored by using the biorthogonal wavelet, and the secret data can be protected by hiding what is adopted as the biorthogonal wavelet.

## VI. FUTURE RESEARCH WORKS

In the future, the author will compare the proposed method with conventional data hiding methods such as steganography method.

## ACKNOWLEDGMENT

The author would like to thank Dr. Kaname Seto of former student of Saga University and Dr. Leland M. Jameson of Naval Research Laboratory for their contribution in this study. The author, also, would like to thank Professor Dr. Hiroshi Okumura and Professor Dr. Osamu Fukuda for their valuable discussions.

## REFERENCES

- [1] Tirkel, A., et al., "Electronic Water Mark" Proceedings DICTA 1993, 666-672, 1993.
- [2] Fabien A. P. Peticolas, Ross J. Anderson, Markus G. Kuhn : "Attacks on copyright marking systems", David Aucsmith(ed), Information Hiding, Proceedings, LNCS 1525, Springer-Verlag, ISBN 3-540-65386-4, pp.219-239, 1998.
- [3] H. Keith Melton : "The Ultimate Spybook", Dorling Kindersley Limited, London, 1996.
- [4] Kohei Arai, Kaname Seto, Data Hiding Based on Wavelet Multiresolution Analysis, Journal of Visual Information Society, Vol.22, Suppl.No.1, 229-232, 2002.
- [5] Kohei Arai Kaname Seto, Data Hiding Based on Multiresolution Analysis Using Information Bias by Eigenvalue Expansion, Journal of Visual Information Society, Vol.23, No.8, pp.72-79, 2003.
- [6] Kohei Arai, Patent Application No. : 2004-29933, Digital Watermark Insertion / Extraction Device and Method.
- [7] Kohei Arai, PCT application number: PCT / JP2005 / 13512, coordinate transformation method, data compression and data hiding method using the same, and their devices, 2005.
- [8] Yao Qiuming, Xuan Guorong, Yang Chengyun, Shi Yunquin, Lossless Data Hiding Using Histogram Shifting Method Based on Integer Wavelets, Proceedings of the IWDW 2006: Digital Watermarking pp 323-332 | Cite as, 2006.
- [9] Kohei Arai, Basic Theory of Wavelet Analysis, Morikita Publishing (November 2000).
- [10] Kohei Arai, Leland Jameson, How to use earth observation satellite data by wavelet analysis, Morikita Publishing (July 2001).
- [11] Kohei Arai, Self-study wavelet analysis, published by Modern Science Co., Ltd. (June 2006).
- [12] Kohei Arai, Method for data hiding based on Legall 5/2 (Cohen-Daubechies-Feauveau: CDF 5/3) wavelet with data compression and random scanning of secret imagery data, International Journal of Wavelets Multi Solution and Information Processing, 11, 4, 1-18, B60006 World Scientific Publishing Company, DOI: 10.1142/SO219691313600060, 1360006-1, 2013.
- [13] Kohei Arai and Yuji Yamada, Improvement of secret image invisibility in circulation image with Dyadic wavelet based data hiding with run-length coding, International Journal of Advanced Computer Science and Applications, 2, 7, 33-40, 2011.
- [14] Cahya Rahmed Kohei Arai, Arief Prasetyo, Noriza Arigki, Noble Method for Data Hiding using Steganography Discrete Wavelet Transformation and Cryptography Triple Data Encryption Standard: DES, IJACSA, 9, 11, 261-266, 2018.
- [15] Kohei Arai, Data Hiding Method Replacing LSB of Hidden Portion for Secrete Image with Run-Length Coded Image, International Journal of Advanced Research on Artificial Intelligence, 5, 12, 8-16, 2016.
- [16] Kohei Arai, Data Hiding Method with Principal Component Analysis and Image Coordinate Conversion, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 12, No. 8, 25-30, 2021.

AUTHORS' PROFILE

Kohei Arai, He received BS, MS and PhD degrees in 1972, 1974 and 1982, respectively. He was with The Institute for Industrial Science and Technology of the University of Tokyo from April 1974 to December 1978 also was with National Space Development Agency of Japan from January, 1979 to March, 1990. During from 1985 to 1987, he was with Canada Centre for Remote Sensing as a Post Doctoral Fellow of National Science and Engineering Research Council of Canada. He moved to Saga University as a Professor in Department of Information Science on April 1990. He was a councilor for the Aeronautics and Space related to the Technology Committee of the Ministry of Science and Technology during from 1998 to 2000. He was

a councilor of Saga University for 2002 and 2003. He also was an executive councilor for the Remote Sensing Society of Japan for 2003 to 2005. He is a Science Council of Japan Special Member since 2012. He is an Adjunct Professor of University of Arizona, USA since 1998. He also is Vice Chairman of the Science Commission "A" of ICSU/COSPAR since 2008 then he is now award committee member of ICSU/COSPAR. He wrote 55 books and published 620 journal papers as well as 450 conference papers. He received 66 of awards including ICSU/COSPAR Vikram Sarabhai Medal in 2016, and Science award of Ministry of Education of Japan in 2015. He is now Editor-in-Chief of IJACSA and IJISA. <http://teagis.ip.is.saga-u.ac.jp/index.html>.



# Multiple Eye Disease Detection using Hybrid Adaptive Mutation Swarm Optimization and RNN

P. Glaret Subin<sup>1</sup>, P. Muthu Kannan<sup>2</sup>

Department of ECE, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences  
Chennai-602105, India.<sup>1,2</sup>

Department of ECE, College of Engineering and Technology, SRM Institute of Science and Technology, Vadapalani Campus  
Chennai- 600026, India<sup>1</sup>

**Abstract**—The major cause of visual impairment in aged people is due to age related eye diseases such as cataract, diabetic retinopathy, and glaucoma. Early detection of eye diseases is necessary for better diagnosis. This paper concentrates on the early identification of various eye disorders such as cataract, diabetic retinopathy, and glaucoma from retinal fundus images. The proposed method focuses on the automated early detection of multiple diseases using hybrid adaptive mutation swarm optimization and regression neural networks (AED-HSR). In the proposed work, the input images are preprocessed and then multiple features such as entropy, mean, color, intensity, standard deviation, and statistics are extracted from the collected data. The extracted features are segmented by using an adaptive mutation swarm optimization (AMSO) algorithm to segment the disease sector from the fundus image. Finally, the features collected are fed to a regression neural network (RNN) classifier to classify each fundus image as normal or abnormal. If the classifier output is abnormal, then it is classified by the corresponding diseases in terms of cataract, glaucoma, and diabetic retinopathy, which improves the accuracy of detection and classification. Ultimately, the results of the classifiers are evaluated by several performance analyses and the viability of structural and functional features is considered. The proposed system predicts the type of the disease with an accuracy of 0.9808, specificity of 0.9934, sensitivity of 0.9803 and F1 score of 0.9861 respectively.

**Keywords**—Adaptive mutation swarm optimization; fundus image; feature extraction; RNN classifier; standard deviation

## I. INTRODUCTION

Nowadays, aged people are mostly affected by chronic diseases such as cataract, glaucoma and diabetic retinopathy, which lead to visual impairment. The optic nerve is damaged due to glaucoma, which results in loss of vision. Glaucoma occurs due to a slow rise in the normal fluid pressure inside the eyes. Cataract occurs due to the clouding of the eye's lens. The progressive damage in the retina's blood vessels, which are essential for good vision of the eye, leads to Diabetic Retinopathy [1]. Based on the supervised learning method, blood vessels are segmented from the fundus image that can be done using Zernike moment-based Shape descriptors and training can be performed using an ANN-based binary classifier to predict cardio vascular diseases [2]. Multiple instances A learning technique is used to classify the diseased image and healthy image, in which the classification can be done by binary classification [3]. Microaneurysms can be recognized using principal component analysis, morphological

processing, averaging filter, and support vector machine classifier. Diabetic Retinopathy disease can also be identified [4]. The early signs of diabetic retinopathy can be identified by applying nineteen features extracted from the fundus image to an artificial neural network, which is trained by Levenberg-Marquardt, and the disease is classified by using Bayesian Regularization [5]. Red lesions can be detected in the blood vessels by using a Gaussian filter and the disease can be predicted using an SVM classifier [6]. Based on the singular value decomposition algorithm, dictionary learning methods can be used to classify healthy people from diabetic patients based on singular Value Decomposition Algorithm [7]. The fundus image is segmented by using a Deep Convolution Neural Network and it increases the accuracy and efficiency in predicting non-proliferated diabetic retinopathy [8]. The blood vessels can be segmented by using dilated convolution, which leads to more accurate detection of ophthalmologic diseases [9]. The two filtering methods, namely median filtering and Gaussian derivative filtering, are used to define the bifurcation point of a blood vessel image segment [10]. DR (Diabetic Retinopathy) can be recognized using the ANN classifier and region growing segmentation to extract exudates, optic plate and veins from the fundus images [11]. DR can be detected by using a reformed capsule network, which attains an accuracy of 97.98% [12]. A hierarchical severe grading system model was developed to detect and classify the different grades of DR. The classifier accuracy is 94% [13]. The optic disc and optic cup boundary of the fundus images are segmented and by using Weighted Least Square fit, holistic features and disc ratio are extracted, and then they are fed to a Convolutional Multi-Layer Neural Network Classifier to classify the glaucoma [14]. A classification method of multi feature analysis along with a Discrete Wavelet transform is used to detect glaucoma. This model classifies glaucoma with an accuracy of 95% [15]. The input fundus image is validated using Le-Net architecture and the optic disc and optic cup are segmented using U-Net Architecture. Glaucoma can be detected with the use of SVM Classifier, Neural Network Classifier, and Adaboost Classifiers [16]. The eyeball area is extracted from the fundus image using an object detection network and multi task learning is applied to detect the cataract [17]. A Deep Convolution Neural Network with Resnet for classification can be used to identify cataract. The systems show an accuracy of 95.77% [18]. Gray Level Co-occurrence Matrix is utilized for feature extraction, and the classification of different levels of cataract can be done by Back Propagation Neural Network Classifier. This system



provides an accuracy of 82.4% [19]. The above-mentioned techniques are utilized to anticipate a single disease from the fundus image of the retina. The proposed paper uses hybrid adaptive mutation swarm optimization and regression neural network (AED-HSR) to provide automated early detection of multiple diseases.

The contributions of the proposed work are:

- Multiple features are extracted from the collected data and standard deviation, smoothness, entropy, shape, color, intensity and statistics are included for feature extraction.
- An adaptive mutation swarm optimization (AMSO) algorithm is used to segment the disease sector from fundus image.
- The collected features are fed to the regression neural network-based classifier to classify each fundus image as normal or abnormal.

The rest of this paper provides the recent related works under Section II, Problem methodology and System model in Section III, the proposed AED-HSR technique using AMSO algorithm and RNN algorithm experimental setup and results are explained in Section IV, and Section V describes the conclusion of AED-HSR.

## II. RELATED WORKS

Kangrok et al., proposed a strategy to detect DR by utilizing automatic segmentation of the ETDRS 7SF in order to expel the undesirable components in the fundus image, and then it was fed to a ResNet – 34 model for the classification of the disease, which provided an accuracy of 83.38% [20]. Saeid et al., proposed a combinational approach of fuzzy C-means and genetic algorithms for the prediction of DR from the angiographic images of diabetic patients, which provided a sensitivity of 78% [21]. Zhuang et al., introduced a weighted voting algorithm to categorize the DR disease and the trained network model was applied to the hospital data, which provided 92% accuracy [22]. Rego et al., worked on a Convolutional Neural Network (CNN) model with Inception-V3 for DR screening of fundus images. In this approach, the model analyzed 295 images and the results were compared with a team of ophthalmologists. This model predicted DR with an accuracy of 95% [23]. Mohammed Hasan et al., suggested a combined method of Convolution Neural Network and Principal Component Analysis for the diagnosis of DR with an accuracy of 98.44% [24]. Hemelings et al., suggested a method based on a deep learning approach to identify glaucoma in which the fundus image was cropped with radius as image size percentage, optic nerve head (ONH) centered with spacing of 10–60%. This model resulted in an AUC of 0.94 [25]. Salam et al., developed an algorithm for glaucoma diagnosis based on combined structural and non-structural features. This method was evaluated with 100 patients' fundus images, which provided a 100% sensitivity and an 87% specificity [26]. Nataraj et al., suggested a machine learning based classification technique to identify glaucoma from fundus images. This method used a unique template approach for segmentation, the Gray Level Coherence Matrix approach

for feature extraction, and wavelet transform for texture and structure-based features to improve the efficiency of the system [27]. Latif et al., proposed a model for detecting glaucoma that had two parts: one to find the optic discs and the other to use transfer learning to find glaucoma. This method provided 95.75% accuracy, 94.75% sensitivity, and 94.90% specificity [28]. Xu et al., proposed a method based on the transfer induced attention network for glaucoma diagnosis that extracted the deep patterns related to disease with limited supervision. The model was evaluated on clinical datasets, which provided an accuracy of 85.7% [29]. A novel method was developed by Raja et al., to identify glaucoma at an earlier stage by using a deep learning approach for segmenting the optic cup and optic disc and an SVM classifier to predict the disease with 92% accuracy [30]. Hasan et al., suggested a convolution neural network to diagnose the cataract disease from the fundus image. This model predicted the cataract disease with an accuracy of 98.17% [31]. Azhar et al., employed CNN for extracting the features and Support Vector Machine (SVM) for the prediction of cataracts. The system model provided an accuracy of 95.65% [32]. Pratap et al., suggested a technique for automatic cataract detection by utilizing singular value decomposition as a feature extractor and SVM as a classifier. The accuracy of the method was 97.78% [33]. Imran et al., developed a strategy for the identification of cataracts. The fundus images were preprocessed and then, by using the combination of Self Organizing Maps and Radial Basis Function (RBF) Neural Network, the model predicted the cataract with an accuracy of 95.3% [34]. Behera et al., [35] used an RBF-based SVM Classifier to predict cataract diseases from fundus images.

## III. METHODOLOGY

The proposed automated early detection of multiple eye diseases by means of hybrid adaptive mutation swarm optimization and regression neural network (AED-HSR) techniques has been used to improve the accuracy of eye disease diagnosis. The proposed system consists of two phases. In the first phase, AMSO is used to segment the blood vessels, optic distance, exudates, and hemorrhage from the extracted features of the preprocessed fundus image. In the second phase, an RNN classifier is utilized to identify multiple eye diseases, namely cataract, diabetic retinopathy, and glaucoma. The flow diagram for the proposed system model for multiple disease detection is shown in Fig. 1.

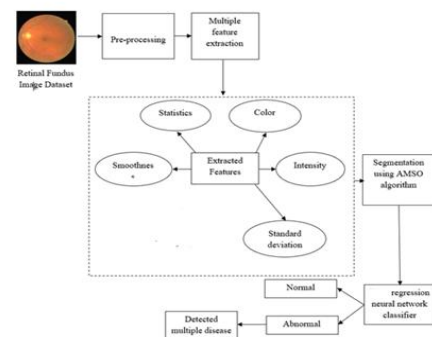


Fig. 1. System Model of Proposed AED-HSR Technique.

### A. Preprocessing

The creation of a binary mask is the initial step in image preprocessing, which is used to recognize that the pixels belong to the Region of Interest. Creating a binary mask avoids unwanted processing of pixels outside the ROI, which will reduce the processing time of an image. Masking can be done by convoluting the red color channel with a Gaussian low pass filter. Then the image undergoes thresholding by using Otsu's global thresholding [36]. The image is then converted to a grayscale image, which has better contrast than the other color channels, such as the red color channel and blue color channel. Since various eye diseases can be identified in the blood vessels of the fundus image, and the contrast between the blood vessels can be easily identified in the green channel of the color image. After the conversion of the grayscale image, the image is resized to standard form because of the wider size of the input image. The resized image is then denoised using a gray level morphological operation to remove the brightness strip located at the center of the blood vessels.

### B. Feature Extraction

Feature extraction is a method of mapping the primary feature into a low dimensional feature space for better classification. Each feature should have a larger variance to distinguish the features from the image. To detect the diseased image, the features such as blood vessels, optic disc, hemorrhage, and retinal exudates should be extracted from the fundus image of the retina. Glaucoma can be identified by the variation in the optic cup, which is a portion of the optic disc. In order to accomplish the automatic diagnosis of glaucoma, the location of the optic disc, which contains more information for glaucoma detection, must be extracted from the fundus image [37]. Optic disc feature extraction can be done by the entropy method. Blood vessels are extracted by using Gabor filters [38].

1) *Entropy*: The average quantity of information owing to the variance of pixel values in an image is defined as entropy in image processing. The image entropy of a distinct brightness values can be calculated by "1"

$$H(I) = \sum_{k=1}^{B_g} P(I_k) \log_2 \left( \frac{1}{P(I_k)} \right) \quad (1)$$

where  $P(I_k)$  denotes the  $k$  brightness value distribution of image  $I$  and  $B_g$  represents the number of brightness levels in an image.

Based on the values of entropy, texture analysis of an image can be done. Lower values of entropy result in the smoothening of texture, whereas higher values of entropy give texture with more details. In the proposed model, higher values of entropy in the fundus image are taken as the optic disc location as it has more details such as nerves and blood vessels.

2) *Gabor filters*: Gabor filters are mostly used to enhance the blood vessels from the fundus image. The product of Gaussian envelope function and complex trigonometric function results in complex Gabor function. To enhance the blood vessels in the fundus image, real portion of the complex Gabor function is utilized and is given by "2"

$$S(u,v,\lambda,\phi,\sigma,\tau) = \exp\left(\frac{-u^2 - v^2 - \tau^2}{2\sigma^2}\right) \cdot \cos(2\pi u' / \lambda + \phi) \quad (2)$$

where  $s$  represents the two-dimensional Gabor kernel function with variables  $u$  and  $v$  and  $u' = u \cos \phi + v \sin \phi$  and  $v' = -u \sin \phi + v \cos \phi$ . Scale ( $\sigma$ ), Wavelength ( $\lambda$ ), orientation ( $\phi$ ) and aspect ratio ( $\tau$ ) are the four parameters to control the shape. The Gabor kernel is rotated at an angle of 15 degree so that 12 different kernels are obtained; it is convolved with the preprocessed image and it selects the utmost response for each pixel. Subsequently, the pixels having blood vessels are more prevailing than the other pixels.

The statistical features, namely mean and standard deviation can be calculated using the following equations "3" and "4"

$$\mu = \sum_{k=0}^{L-1} k p(k) \quad (3)$$

$$\sigma^2 = \sum_{k=0}^{L-1} (k - \mu)^2 p(k) \quad (4)$$

### C. Proposed AED-HSR Technique using AMSO and RNN Algorithm

In this section, AMSO is described in Section.4.1 and Regression neural network classifier to identify the diseased image is explained briefly in Section.4.2.

1) *Segmentation using AMSO algorithm*: The features retrieved from the fundus image are segmented using the Adaptive Mutation Swarm Optimization (AMSO) technique. Basically, Swarm will be initiated by the Particle Swarm Optimization. The solution of each search space is determined based on the position and the velocity of the particle in the swarm. The position and the velocity of the particle are changed for each iteration and the global best position  $p_g$  and personal best position  $p_i$  are found out. To reach the ideal PSO state, a larger interference amplitude is needed in the early iteration phase to ensure better global search capabilities and a smaller interference amplitude is required in the late iteration phase to ensure convergence.

a) *Chase-Swarming Behavior*: where  $x_i$  represents the position of cockroach, step denotes a fixed value, 'rand' represents a random number lying between 0 and 1,  $p_i$  is the personal best position, and  $p_g$  is the global best position. The personal best position of cockroach of size  $N$  is given by "5" and "6"

$$X_i = x_i + \text{step} \cdot \text{rand} \cdot (p_i - x_i), x_i \neq p_i \quad (5)$$

$$X_i = x_i + \text{step} \cdot \text{rand} \cdot (p_g - x_i), x_i = p_i \quad (6)$$

where visual is a constant,  $j = 1, 2, \dots, N$ ,  $i = 1, 2, \dots, N$ . The global best position of the cockroach is given by "7" and "8"

$$p_i = \text{opt}_j \{x_j | |x_i - x_j| \leq \text{Visual}\} \quad (7)$$

$$p_g = \text{opt} \{x_i\} \quad (8)$$

The inertial weight to chase swarming component of original AMSO is given by "9" and "10"

$$X_i = \omega \cdot x_i + \text{step} \cdot \text{rand} \cdot (p_i - x_i), x_i \neq p_i \quad (9)$$

$$X_i = \omega \cdot x_i + \text{step.rand.}(p_g - x_i), x_i = p_i \quad (10)$$

where  $\omega$  is an inertial weight of the particle.

*b) Hunger Behavior:* In this paper, an enhanced cockroach swarm optimization is prolonged with an extra feature called hunger behavior. For a particular period of time, when the cockroach is hungry, it migrates from its original position and searches for food source  $x_{\text{food}}$  within the search space. Partial differential equation (PDE) migration technique is used for hunger behaviour modelling. Hunger behaviour impedes local optimization and increases population diversity.

Kerckhove defines the PDE migration equation as

$$\partial p / \partial t = -s \partial p / \partial x \quad (11)$$

with initial population distribution  $p(0, x) = p_0(x)$ .

where the parameter  $s$  controls the migration speed.  $p$  represents population size,  $t$  denotes the time, and  $x$  represents the location or position.  $(t, x)$  is the population size at time  $t$  in location  $x$

Equation (9) can be expressed as

$$\partial p / \partial t + s \partial p / \partial x = 0 \quad (12)$$

By integration, we have

$$x - st = \sigma, p = p(\beta), p = p(x - st)$$

$$p[t, x] = p_0[x - st]$$

Since displacement is the product of speed and time, in  $p_0(x - st)$ ,  $p_0(x)$  is replaced by  $st$ .

$p_0(x - st)$ , satisfies PDE at any initial population distribution  $p_0(x)$ . Hunger behavior is defined as follows: If hunger is equal to threshold hunger  $t_{\text{hunger}}$ , then the new position of cockroach is

$$X_i = x_i + (x_i - st) + x_{\text{food}} \quad (13)$$

where  $x_i$  denotes the position of the cockroach,  $(x_i - st)$  represents the migrated position of the cockroach from its current position, and hunger is a random number which lies in the range of 0 to 1.

#### D. Classifying Feature Models using RNN

The segmented portions of the fundus images are fed into the Regression Neural Network Classifier for the prediction of diseases in the form of a regression task. The architecture of the RNN is shown in Fig. 2. The segmented features are given as input to the RNN, which contains a fully connected layer and batch normalization with a Leaky Rectified Linear Unit (FC-BN-LReLU), pursued by a dropout layer with a dropping probability of 0.2 to avoid the overfitting problems, and two layers of FC-BN-LReLU, followed by a drop out layer. The following layer is a fully connected layer with LReLU, which does not contain batch normalization. The last layer is a fully connected layer without any activation function that will classify the type of disease.

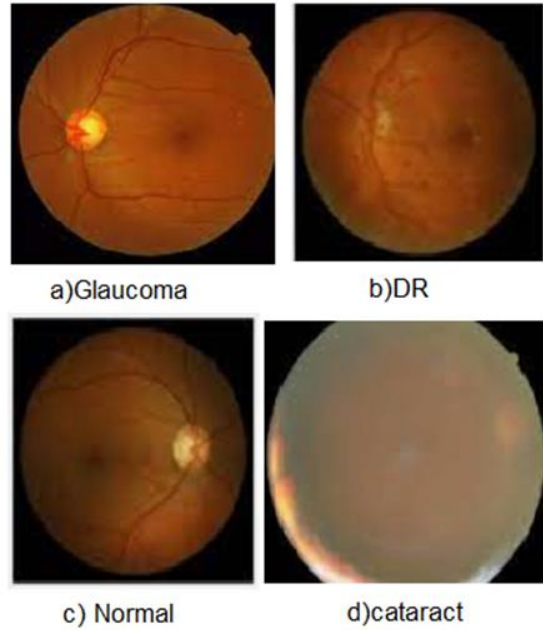


Fig. 2. Test Image Samples.

#### IV. EXPERIMENTAL SET-UP AND RESULTS

In the proposed AED-HSR, Cataract, glaucoma, DR and normal images are classified from the fundus images and the performance analysis is calculated based on distinct algorithms. The proposed method was implemented in the online datasets using MatlabR2022a.

##### A. Dataset

The proposed AED-HSR method was implemented on the dataset that comprised of 800 normal images, 800 cataract images, 800 diabetic retinopathy images, and 800 glaucoma images, taken from the Ocular Disease Intelligent Recognition (ODIR) database, which contains the ophthalmic database of 5000 patients, including their ages, right and left eye fundus images, and the doctor's diagnostic keywords. The images are collected by the Shanggong Medical Technology Co., Ltd. from numerous hospitals in China.

##### B. Results

In the experimental set-up, 30% of images are utilized for testing and 70% of images are utilized for training. The test image samples are shown in Fig. 2. A learning rate of 0.01 with a maximum of 39 epochs was used in the training phase of the proposed system. The proposed method utilizes 5 distinct features such as smoothness, statistics, color, intensity, and standard deviation to train the RNN. The training performance of the AED-HSR in terms of training, validation and test data is shown in Fig. 3. From the training progress of AED-HSR, the mean square error (MSE) drops rapidly in the first 10 epochs and the best validation performance is 0.074662 at epoch 33. The training MSE gradually drops a bit and the stability is more in the final epochs. In the training model, more attributes are used to aid in prediction which may be useful to prevent overfitting.

The training state of the proposed AED-HSR shown in Fig. 4 will provide the gradient of 0.16819 at epoch 39 and the maximum mu value of 0.001 is reached at epoch 39 which controls the neurons weight updating process during training. By applying the test images as an input to the model, AED-HSR classifies the image as a diseased image or normal image.

A regression analysis was performed in order to determine the relationship between the network output and the corresponding targets. The Regression plot of the AED -HSR model is shown in Fig. 5 which shows a correlation between the two sets of data. It demonstrates that there is a good fit between the values that were predicted by AED-HSR and the actual measured data with higher values of R. The regression plot indicates that the outputs closely match the targets with an R-Value. The Confusion matrix for the four different models is shown in Fig. 6. From the values of Confusion Matrix (VCM), True Positive (TP), True Negative, False Positive and False Negative values are calculated in order to analyze the performance of the system model. Table I shows the prototype of VCM. Table II shows the characteristics of VCM.

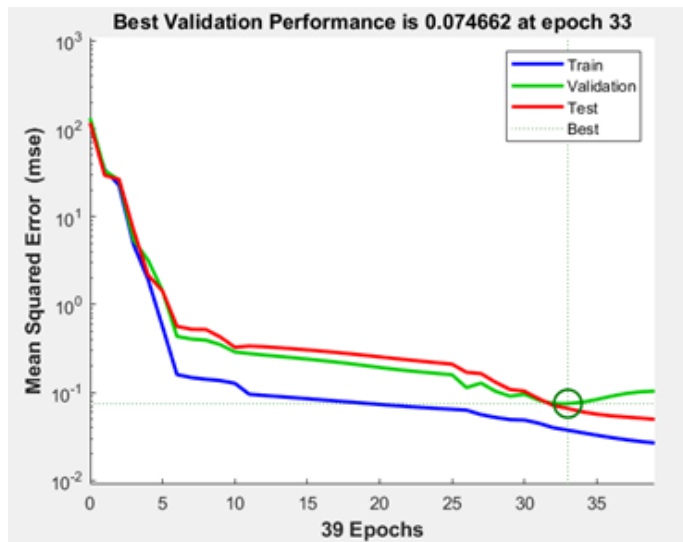


Fig. 3. Training Progress of the AED-HSR Model.

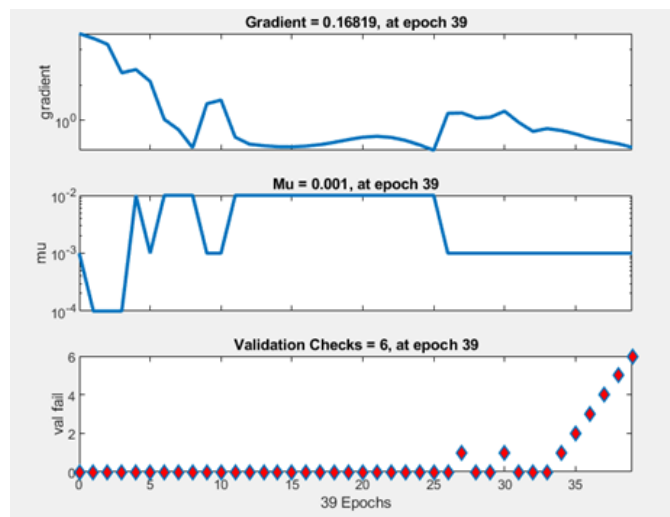


Fig. 4. Training State of the Proposed AED-HSR Model.

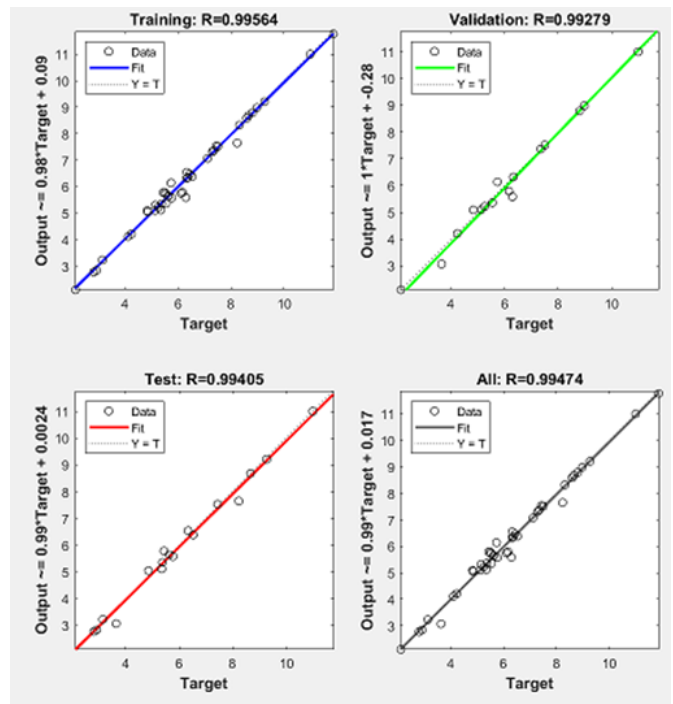


Fig. 5. Regression Plot of the Proposed AMSO-RNN Model.

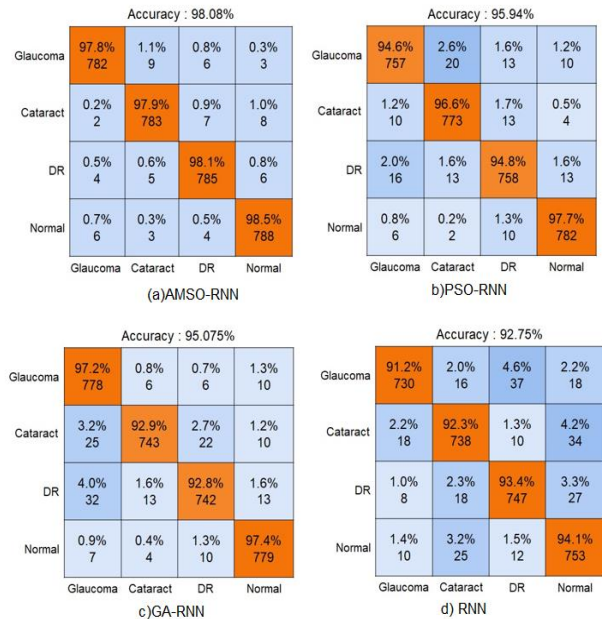


Fig. 6. Confusion Matrix of Four Different Methods.

TABLE I. VCM PROTOTYPE

		PREDICTED			
		X	Y	Z	R
ACTUAL	X	$L_{XX}$	$L_{XY}$	$L_{XZ}$	$L_{XR}$
	Y	$L_{YX}$	$L_{YY}$	$L_{YZ}$	$L_{YR}$
	Z	$L_{ZX}$	$L_{ZY}$	$L_{ZZ}$	$L_{ZR}$
	R	$L_{RX}$	$L_{RY}$	$L_{RZ}$	$L_{RR}$



TABLE II. CHARACTERISTICS OF VCM

<b>TRUE POSITIVE</b>	$TP_X = L_{XX}$
	$TP_Y = L_{YY}$
	$TP_Z = L_{ZZ}$
	$TP_R = L_{RR}$
<b>FALSE POSITIVE</b>	$FP_X = L_{YX} + L_{ZX} + L_{RX}$
	$FP_Y = L_{XY} + L_{ZY} + L_{RY}$
	$FP_Z = L_{XZ} + L_{YZ} + L_{RZ}$
	$FP_R = L_{XR} + L_{YR} + L_{ZR}$
<b>TRUE NEGATIVE</b>	$TN_X = L_{YZ} + L_{YY} + L_{ZY} + L_{YR} + L_{ZZ} + L_{RY} + L_{RZ} + L_{YR} + L_{RR}$
	$TN_Y = L_{XX} + L_{ZZ} + L_{XZ} + L_{XR} + L_{ZX} + L_{RR} + L_{RX} + L_{RZ} + L_{ZR}$
	$TN_Z = L_{RY} + L_{RR} + L_{XX} + L_{XY} + L_{XR} + L_{YX} + L_{YY} + L_{YR} + L_{RX}$
	$TN_R = L_{ZY} + L_{ZZ} + L_{XX} + L_{XY} + L_{XR} + L_{YX} + L_{YY} + L_{YZ} + L_{ZX}$
<b>FALSE NEGATIVE</b>	$FN_X = L_{XY} + L_{XZ} + L_{XR}$
	$FN_Y = L_{YX} + L_{YZ} + L_{YR}$
	$FN_Z = L_{ZX} + L_{XY} + L_{ZR}$
	$FN_R = L_{RX} + L_{RY} + L_{RZ}$

The proposed AED- HSR model detects the eye disease with an accuracy of 98.08%. DR detection from the proposed model is shown in Fig. 7. The classification of glaucoma using the proposed system model is shown in Fig. 8. Cataract image detection by utilizing the proposed AED-HSR model is shown in Fig. 9.

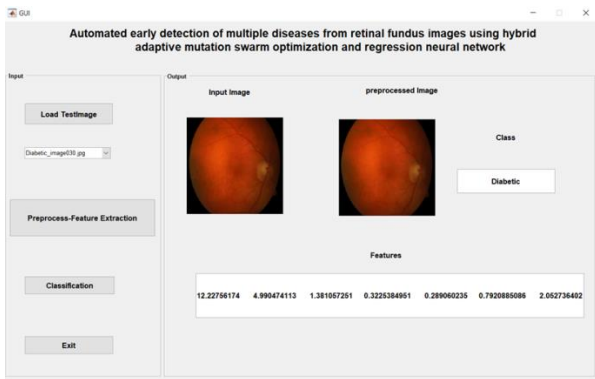


Fig. 7. Detection of DR.

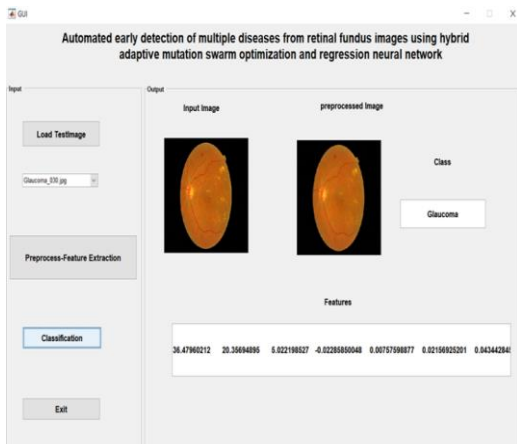


Fig. 8. Detection of Glaucoma.

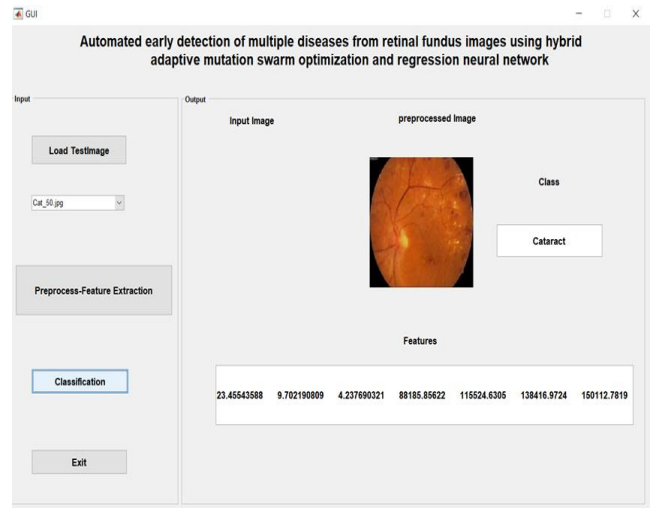


Fig. 9. Detection of Cataract.

C. Performance Analysis

The proposed model performance can be analyzed in terms of accuracy, sensitivity, specificity, negative predictive value (NPV), positive predictive value (PPV), false negative rate (FNR), false positive rate (FPR), and false discovery rate (FDR) from VCM. Table III shows the parameters of the performance metrics. Sensitivity alludes to a test’s capacity to assign a diseased image as positive. Specificity characterizes the capacity of the test to assign the non-diseased image as negative.

Accuracy is defined as the ratio of correctly predicted images to the total images. PPV refers to the probability that the individual has the disease when confined to those individuals who test positive. NPV measures the proportion of genuine negative expectations considering all negative forecasts. FNR is calculated by the ratio of false negative to total positive. The FDR measures the extent of the alerts that are irrelevant. Fig. 10 shows the performance metrics for each class of eye diseases based on RNN.

Fig. 10 shows that RNN has 94.1% accuracy rate for identifying the normal eyes. The maximum specificity for glaucoma affected eye is 98.5%. F1 score for DR is 95.41%.

TABLE III. PERFORMANCE METRICS PARAMETERS

Parameters	Formula
Sensitivity	$Sen = TP / (TP + FN)$
Specificity	$Spec = TN / (TN + FP)$ ;
Accuracy	$Acc = (TP+TN) / (TP+TN+FP+FN)$
False positive rate	$FPR = 1 - Spec$
False negative rate	$FNR = FN / (TP + FN)$
Positive predictive value	$PPV = TP / (TP + FP)$
Negative predictive value	$NPV = TN / (TN + FN)$ $NPV = TN / (TN + FN)$
False discovery rate	$FDR = 1 - PPV$
F1_Score	$F1\_Score = (2*TP) / ((2*TP) + FP + FN)$

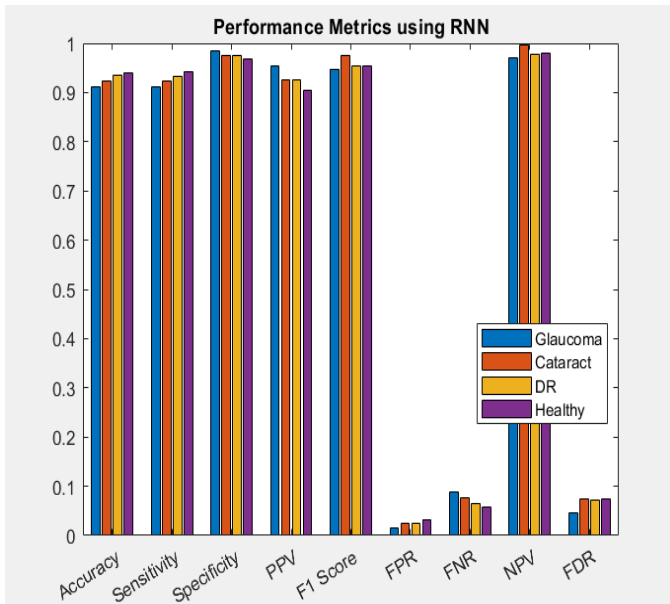


Fig. 10. Performance Metrics of each Class using RNN.

Fig. 11 indicates the performance metrics for each class of eye diseases based on GA-RNN. Fig.11 shows that GA-RNN has 97.4% accuracy rate for identifying the normal eyes. The maximum specificity for cataract affected eye is 99.04%. F1 score for Normal eye is 97.99%.

Fig. 12 indicates the performance metrics for each class of eye diseases based on PSO-RNN. Fig.13 shows that PSO-RNN has 97.7% accuracy rate for identifying the normal eyes. The maximum specificity for Normal eye is 98.87%. F1 score for Normal eye is 98.30%.

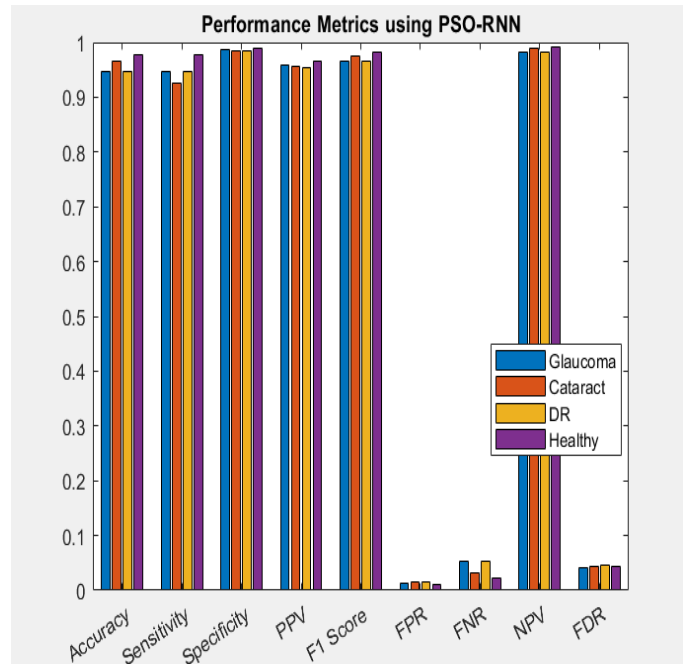


Fig. 12. Performance Metrics of each Class using PSO-RNN.

Fig. 13 indicates the performance metrics for each class of eye diseases based on the proposed AMSO-RNN. Fig. 13 shows that AMSO-RNN has 98.5 % accuracy rate for identifying the normal eyes. The maximum specificity of Glaucoma diseased eye is 99.5% F1 score for Normal eye is 98.83%.

Overall, the system automatically detects the eye disease with 98.08 % accuracy, 99.34% specificity, 98.03% sensitivity, 98.03% PPV, 99.34% NPV, 0.62% FPR, 1.93% FNR, 98.67% F1 Score and 1.96% FDR.

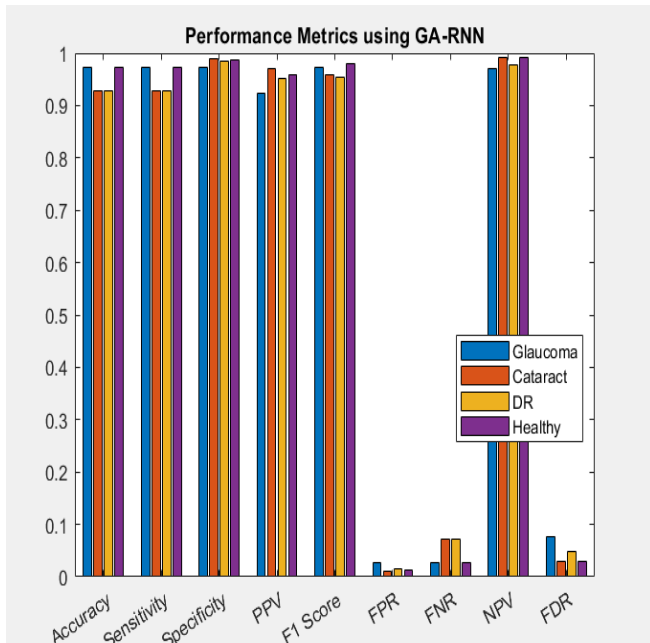


Fig. 11. Performance Metrics of each Class using GA- RNN.

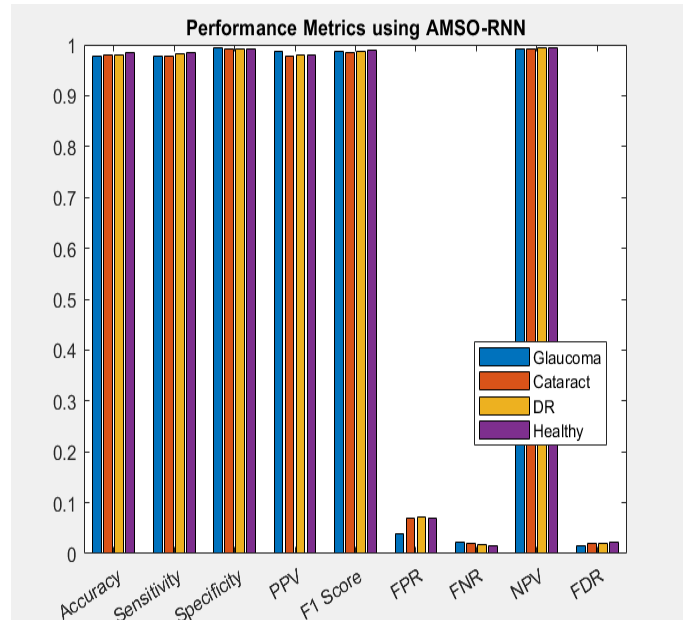


Fig. 13. Performance Metrics of each Class using AMSO-RNN.



The performance Matrix of AMSO-RNN is compared with that of the other strategies, namely RNN-PSO, RNN-GA, and RNN, as shown in the Fig. 14. From the comparison measures of four different methods, the proposed method has an improvement of 2.23% in accuracy, 2.18% in sensitivity, 0.71% in specificity, 2.19% increase in PPV, 1.45% increase in F1 Score, 53.38% decrease in FPR, 51.99% decrease in FNR, 0.72% increase in NPV and 54.2% decrease in FDR when compared with RNN-PSO. Also, the proposed AMSO-RNN provides an improvement of 3.16 % in accuracy, 3.14% in sensitivity, 1.01% in specificity, 3.08 % increase in PPV, 2.09% increase in F1 Score, 61.73% decrease in FPR, 60.61% decrease in FNR, 1.01% increase in NPV and 58.69% decrease in FDR when compared with RNN-GA and an improvement of 5.75% in accuracy, 5.72% in sensitivity, 1.88% in specificity, 5.68% increase in PPV, 3.79% increase in F1 Score, 74.05% decrease in FPR, 73.27% decrease in FNR, 1.83% increase in NPV and 70.61% decrease in FDR when compared with RNN.

**D. Comparison of Proposed Method with Previous Studies**

This subsection details the performance analysis comparison of the proposed method with other state of art methods. Table IV shows the performance comparison of the proposed method with previous strategies.

TABLE IV. COMPARISON OF PROPOSED METHOD WITH PREVIOUS STUDIES

Author	Task	Method	Accuracy (%)	Sensitivity (%)	Specificity (%)	F-1 Score (%)
Oh et al. [20]	DR	ETDRS&SF	83.38	80.6	83.41	-
Latif et al. [28]	Glaucoma	ODG-NET	95.75	94.90	94.75	-
AZhar et al. [32]	Cataract	CNN - SVM	95.6	-	-	-
Tayal et al [39]	DME, Drusen, Choroidal, Normal	DL-CNN	96.5	-	98.6	-
Proposed method	Glaucoma DR, Cataract, Normal	AMSO-RNN	98.08	98.03	99.34	98.67

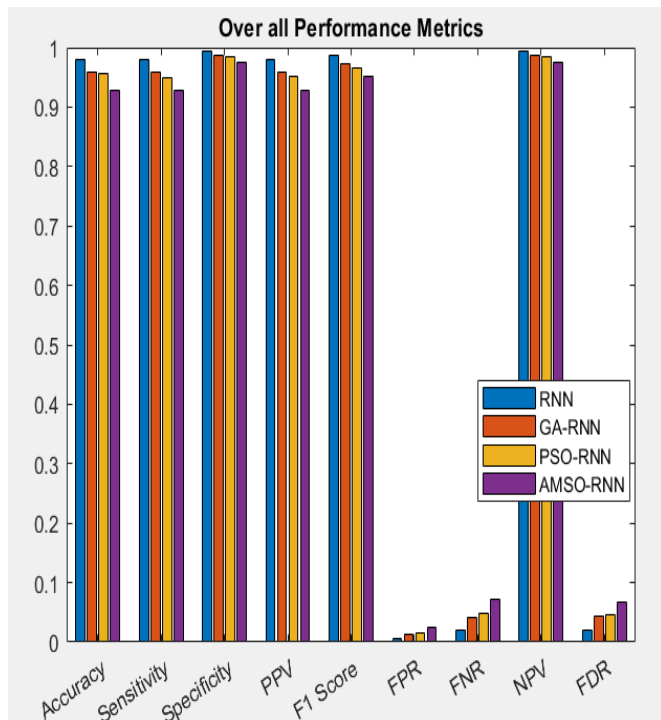


Fig. 14. Overall Performance Metrics.

**V. CONCLUSION**

From retinal fundus images, the proposed AED-HSR technique automatically detects the types of different eye diseases such as DR, glaucoma, and cataract from the retinal fundus images. In this approach, the input images are preprocessed by masking, thresholding, and resizing. The resized image is denoised by using gray level morphological transformation. The preprocessed image is then subjected to feature extraction in order to retrieve the image's statistical features. The collected features are then segmented using AMSO and given to the RNN Classifier. The proposed algorithm has been tested on an ODIR database from the Kaggle datasets. The proposed system predicts the type of the disease with 98.08% accuracy, 99.34% specificity, 98.03% sensitivity, 98.03% PPV, 99.34% NPV, 0.62% FPR, 1.93% FNR, 98.67% F1 Score and 1.96% FDR. The proposed methods provide better results in contrast with the other techniques such as RNN-PSO, RNN-GA, and RNN. The model that has been proposed provides better performance metrics when compared with the other networks in terms of accuracy, specificity, sensitivity, PPV, NPV, FPR, FNR, FNR, F1Score and FDR. The outcome of this study points to the enhancement of the suggested network architecture as work that should be done in order to achieve future improvements in terms of performance.

REFERENCES

- [1] Centers for disease control and prevention: <https://www.cdc.gov/visionhealth/basics/ced/index.html>.
- [2] Adapa D, Joseph Raj AN, Aliseti SN, Zhuang Z, K. G, Naik G, (2020) .A supervised blood vessel segmentation technique for digital Fundus images using Zernike Moment based features.PLoS ONE 15(3),[tps://doi.org/10.1371/journal.pone.0229831](https://doi.org/10.1371/journal.pone.0229831).
- [3] P. Costa, A. Galdran, A. Smailagic and A. Campilho, (2018). A Weakly-Supervised Framework for Interpretable Diabetic Retinopathy Detection on Retinal Images, in IEEE Access, vol. 6, pp. 18747-18758, doi: 10.1109/ACCESS.2018.2816003.
- [4] S. Kumar and B. Kumar, (2018) Diabetic Retinopathy Detection by Extracting Area and Number of Microaneurysm from Colour Fundus Image, 2018 5th International Conference on Signal Processing and Integrated Networks (SPIN), Noida,pp.359-364, doi: 10.1109/SPIN.2018.8474264.
- [5] N. H. Harun, Y. Yusof, F. Hassan and Z. Embong, (2019). Classification of Fundus Images For Diabetic Retinopathy using Artificial Neural Network, 2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT), Amman, Jordan, pp. 498-501, doi: 10.1109/JEEIT.2019.8717479.
- [6] S. Ekatpure and R. Jain, (2018). Red Lesion Detection in Digital Fundus Image Affected by Diabetic Retinopathy,2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), Pune, India, pp. 1-4, doi: 10.1109/ICCUBEA.2018.8697387.
- [7] N. Karami and H. Rabbani (2017). A dictionary learning based method for detection of diabetic retinopathy in color fundus images, 2017 10th Iranian Conference on Machine Vision and Image Processing (MVIP), Isfahan,, pp. 119-122.
- [8] L. Qiao, Y. Zhu and H. Zhou,Diabetic,(2020). Retinopathy Detection Using Prognosis of Microaneurysm and Early Diagnosis System for Non-Proliferative Diabetic Retinopathy Based on Deep Learning Algorithms, in IEEE Access, vol. 8, pp. 104292-104302, doi: 10.1109/ACCESS.2020.2993937.
- [9] P. Lopes, A. Ribeiro and C. A. Silva, (2019). Dilated Convolutions in Retinal Blood Vessels Segmentation, 2019 IEEE 6th Portuguese Meeting on Bioengineering (ENBENG), Lisbon, Portugal, pp. 1-4, doi: 10.1109/ENBENG.2019.8692520.
- [10] E. Sutanty, D. A. Rahayu, Rodiah, D. T. Susetianingti and S. Madenda,( 2017). Retinal blood vessel segmentation and bifurcation detection using combined filters, 2017 3rd International Conference on Science in Information Technology (ICSITech), Bandung, pp. 563-567,doi:10.1109/ICSITech.2017.8257176.
- [11] Preethy Rebecca, P. Allwin, S. (2021). Detection of DR from retinal fundus images using prediction ANN classifier and RG based threshold segmentation for diabetes. J Ambient Intell Human Comput 12, 10733–10740 doi.org/10.1007/s12652-020-02882-3.
- [12] Kalyani, G., Janakiramaiah, B., Karuna, A. et al. (2021).Diabetic retinopathy detection and classification using capsule networks. Complex Intell. Syst. <https://doi.org/10.1007/s40747-021-00318-9>.
- [13] Bhardwaj, C., Jain, S. & Sood, M. (2021). Hierarchical severity grade classification of non-proliferative diabetic retinopathy. J Ambient Intell Human Comput 12, 2649–2670 <https://doi.org/10.1007/s12652-020-02426-9>.
- [14] Mansour, R.F., Al-Marghilnai, A. (2021) .Glaucoma detection using novel perceptron based convolutional multi-layer neural network classification. Multidim Syst Sign Process 32, 1217–1235. <https://doi.org/10.1007/s11045-021-00781-0>.
- [15] Ajesh, F., Ravi, R. & Rajakumar, G. (2021). Early diagnosis of glaucoma using multi-feature analysis and DBN based classification. J Ambient Intell Human Comput 12, 4027–4036. <https://doi.org/10.1007/s12652-020-01771-z>.
- [16] Rutuja Shinde, ,(2021) Glaucoma detection in retinal fundus images using U-Net and supervised machine learning algorithms, Intelligence-Based Medicine, 5 100038, <https://doi.org/10.1016/j.ibmed.2021.100038>.
- [17] H. Wu, J. Lv and J. Wang, (2021)Automatic Cataract Detection with Multi-Task Learning,2021 International Joint Conference on Neural Networks (IJCNN), pp. 1-8, doi: 10.1109/IJCNN52387.2021.9533424.
- [18] M. R. Hossain, S. Afroz, N. Siddique and M. M. Hoque, (2020), Automatic Detection of Eye Cataract using Deep Convolution Neural Networks (DCNNs), 2020 IEEE Region 10 Symposium (TENSYP), pp. 1333-1338.
- [19] T. S. Mulati and F. Utamingrum, (2021) Hidden Neuron Analysis for Detection Cataract Disease Based on Gray Level Co-occurrence Matrix and Back Propagation Neural Network, 2021 International Conference on ICT for Smart Society (ICISS), pp. 1-5, doi: 10.1109/ICISS53185.2021.9533263.
- [20] Oh, K., Kang, H.M., Leem, D. et al. (2021). Early detection of diabetic retinopathy based on deep learning and ultra-wide-field fundus images. Sci Rep 11, 1897 <https://doi.org/10.1038/s41598-021-81539-3>.
- [21] Saeid Jafarzadeh Ghouschi, Ramin Ranjbarzadeh, Amir Hussein Dadkhah, Yaghoob Pourasad, Malika Bendechache, (2021),An Extended Approach to Predict Retinopathy in Diabetic Patients Using the Genetic Algorithm and Fuzzy C-Means, BioMed Research International, , <https://doi.org/10.1155/2021/5597222>.
- [22] Ai Zhuang, Huang Xuan, Fan Yuan, Feng Jing, Zeng Fanxin, Lu Yaping Detection Algorithm of Diabetic Retinopathy Based on Deep Ensemble Learning and Attention Mechanism Frontiers in Neuroinformatics 15, 2021, DOI=10.3389/fninf.2021.778552, ISSN=1662-5196.
- [23] Rego S, Dutra-Medeiros M, Soares F, Monteiro-Soares M, (2021), Screening for Diabetic Retinopathy Using an Automated Diagnostic System Based on Deep Learning: Diagnostic Accuracy Assessment. Ophthalmologica ,pp250-257.
- [24] Mohammedhasan, M., Uğuz, H. (2020),A new early-stage diabetic retinopathy diagnosis model using deep convolutional neural networks and principal component analysis. Traitement du Signal, 37, 711-722. <https://doi.org/10.18280/ts.370503>.
- [25] Hemelings, R., Elen, B., Barbosa-Breda, J. et al. (2021). Deep learning on fundus images detects glaucoma beyond the optic disc. Sci Rep 11, 20313 <https://doi.org/10.1038/s41598-021-99605-1>.
- [26] Salam, A.A., Khalil, T., Akram, M.U. et al. (2016).Automated detection of glaucoma using structural and nonstructural features. Springer Plus 5, 1519 <https://doi.org/10.1186/s40064-016-3175-4>.
- [27] Nataraj Vijapur, & R. Srinivasa Rao Kunte, (2020) . Efficient Machine Learning Techniques to Detect Glaucoma using Structure and Texture based Features. International Journal of Recent Technology and Engineering (IJRTE), 9 ,pp.193–201.
- [28] Latif, J., Tu, S., Xiao, C. et al. (2022). ODGNet: a deep learning model for automated optic disc localization and glaucoma classification using fundus images. SN Appl. Sci. 4, 98 ,<https://doi.org/10.1007/s42452-022-04984-3>.
- [29] Xu, X., Guan, Y., Li, J. et al. (2021),Automatic glaucoma detection based on transfer induced attention network. BioMed Eng OnLine 20, 39. <https://doi.org/10.1186/s12938-021-00877-5>.
- [30] Raja, J., Shanmugam, P. & Pitchai, R. (2021) An Automated Early Detection of Glaucoma using Support Vector Machine Based Visual Geometry Group 19 (VGG-19) Convolutional Neural Network. Wireless Pers Commun 118,pp. 523–534.
- [31] Md Kamrul Hasan, Tanjum Tanha, Md Ruhul Amin, Omar Faruk, Mohammad Monirujjaman Khan, Sultan Aljahdali, Mehedi Masud, (2021),Cataract Disease Detection by Using Transfer Learning-Based Intelligent Methods, Computational and Mathematical Methods in Medicine, Article ID 7666365, 11 pages .
- [32] Azhar Imran, Jianqiang Li, Yan Pei, Faheem Akhtar, Ji-Jiang Yang & Yanping Dang,(2020) . Automated identification of cataract severity using retinal fundus images, Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization, 8:6, pp. 691-698, DOI: 10.1080/21681163.2020.1806733.
- [33] T. Pratap and P. Kokil, (2019).Automatic Cataract Detection in Fundus Retinal Images using Singular Value Decomposition,2019 International Conference on Wireless Communications Signal Processing and Networking (WiSPNET), pp. 373-377, doi: 10.1109/WiSPNET45539.2019.9032867.

- [34] A. Imran, J. Li, Y. Pei, F. Akhtar, J. -J. Yang and Q. Wang, (2019), Cataract Detection and Grading with Retinal Images Using SOM-RBF Neural Network, 2019 IEEE Symposium Series on Computational Intelligence (SSCI), pp. 2626-2632.
- [35] M. K. Behera, S. Chakravarty, A. Gourav and S. Dash, (2020), Detection of Nuclear Cataract in Retinal Fundus Image using Radial Basis Function based SVM, 2020 Sixth International Conference on Parallel, Distributed and Grid Computing (PDGC), pp. 278-281, doi: 10.1109/PDGC50313.2020.9315834.
- [36] F. A. Hashim, N. M. Salem and A. F. Seddik, (2013). Preprocessing of color retinal fundus images, 2013 Second International Japan-Egypt Conference on Electronics, Communications and Computers (JEC-ECC), 6th of October City, pp. 190-193, doi: 10.1109/JEC-ECC.2013.6766410.
- [37] Lamia Abed Noor Muhammed, Localizing Optic Disc in Retinal Image Automatically with Entropy Based Algorithm, International Journal of Biomedical Imaging, (2018) Article ID 2815163, 7 pages, doi.org/10.1155/2018/2815163.
- [38] Zafer Yavuz, Cemal Köse, (2017). Blood Vessel Extraction in Color Retinal Fundus Images with Enhancement Filtering and Unsupervised Classification, Journal of Healthcare Engineering, Article ID 4897258, 12 pages, doi.org/10.1155/2017/4.
- [39] Tayal, A., Gupta, J., Solanki, A. et al. (2021). DL-CNN-based approach with image processing techniques for Diagnosis of retinal diseases. Multimedia Systems. <https://doi.org/10.1007/s00530-021-00769-7>.

# Insect Pest Image Detection and Classification using Deep Learning

Niranjan C Kundur<sup>1</sup>

Assistant Professor

Department of Computer Science and Engineering  
JSS Academy of Technical Education, Bengaluru, India

P B Mallikarjuna<sup>2</sup>

Associate Professor

Department of Computer Science and Engineering  
JSS Academy of Technical Education, Bengaluru, India

**Abstract**—Farmers' primary concern is to reduce crop loss because of pests and diseases, which occur irrespective of the cultivation process used. Worldwide more than 40% of the agricultural output is lost due to plant pathogens, insects, and weed pests. Earlier farmers relied on agricultural experts to detect pests. Recently Deep learning methods have been utilized for insect pest detection to increase agricultural productivity. This paper presents two deep learning models based on Faster R-CNN Efficient Net B4 and Faster R-CNN Efficient Net B7 for accurate insect pest detection and classification. We validated our approach for 5, 10, and 15 class insect pests of the IP102 dataset. The findings illustrate that our proposed Faster R-CNN Efficient Net B7 model achieved an average classification accuracy of 99.00 %, 96.00 %, and 93.00 % for 5, 10, and 15 class insect pests outperforming other existing models. To detect insect pests less computation time is required for our proposed Faster R-CNN method. The investigation reveals that our proposed Faster R-CNN model can be used to identify crop pests resulting in higher agricultural yield and crop protection.

**Keywords**—Deep learning; faster RCNN; insect pest detection; IP102 dataset; efficient net

## I. INTRODUCTION

Agricultural production from field crops has advanced quickly in both quantity and quality, but the prevalence of pests and diseases on crops has limited the quality of agrarian output. If pests on crops are not thoroughly inspected and a sufficient, long-lasting treatment is not offered, the quality and amount of food production will be lowered, causing an increase in poverty and food shortages. Any country's economy might be negatively impacted by this, but it would be most harmful in places where 60-70% of the populace relies completely on income from the agricultural sector to support itself. Getting rid of pests that are growing and reducing crop production is a significant issue for agricultural producers. According to our research, a pest is any species that disperses disease and induces damage to the plants. Aphids, flax budworm, flea beetle, cabbage butterfly, peachtree borer, prodenia litura, thrips and mole cricket are the most frequent pests that attack plants. In order to prevent a large amount of loss and boost crop yields, it is necessary to identify these pests at all phases of their life cycles, whether they are nascent or advanced. Understanding and classifying insects is the initial step in preventing crop damage caused by insect pests. This will allow us to distinguish between harmless insects and dangerous ones. In recent times, there has been a rise in

awareness of automated pests' classification because this activity necessitates ongoing, intensive monitoring [1]. It is commonly known that distinct insect species may have phenotypes that are similar to one another and that due to various habitats and growth cycles, insects can have intricate morphologies [2] [3]. An outstanding method for recognizing insect images has been made possible by the development of machine learning techniques. Vehicle recognition and motion detection have seen considerable success utilizing computer vision as well as machine learning techniques [4] [5]. A sizable pest dataset of 40 high-grade pest categories was labeled using a multi-level classification framework of alignment-pooling method [6]. A dataset with 563 pest images partitioned into 10 categories was used. To classify the dataset, training was done on a Support Vector Machine for custom features [7]. Various image processing techniques to detect and retrieve insect pests by developing a machine-driven detection and removal system for evaluating pest concentration in paddy crops [8]. To identify the pest from a dataset of pest images K mean segmentation technique was implemented. In order to classify the pests, the discrete cosine transform method was implemented and the pest images were classified using an artificial neural network. Images were validated for five pests and obtained an accuracy of 94.00 % [9]. Deep learning techniques like convolutional neural networks have lately been used in agricultural production as a viable approach for fully automated pest classification [10]. The convolutional neural networks exert a significant influence on image elements and has their own feature extractor, which makes them superior to conventional image processing techniques and machine learning. Additionally, in several applications of medical image analysis, convolutional neural networks demonstrated their ability to manage picture noise and illumination change [11]. In this study, a Faster R-CNN framework to detect and classify insect pests is investigated.

The main contributions of this work are as follows:

- 1) To detect and classify crop pests, a Faster R-CNN framework with Efficient Net is used. In order to improve the performance of the model the network drop connects is used to prevent over fitting and to increase regularization effect a swish function is utilized for Efficient Net.
- 2) The Region Proposal Network module and the bounding box regression can accurately predict the classes and

locations of various crop pests. The computational time required for detecting the insect pests is less.

3) Compared to other methods, the evaluation results of insect pest classification using the proposed Faster R-CNN framework demonstrated superior performance.

## II. RELATED WORK

Several deep learning techniques have been used recently to categorize pests and develop cutting-edge outcomes in several applications for pest identification. Convolutional neural network and saliency techniques were used for classifying insect pests. Image processing algorithms known as saliency approaches emphasize the most important areas of an image. These techniques are based on the realization that the observer accurately distinguishes between the portions of its field of vision that are important and those that are not useful, rather than focusing on the entire range of vision. They obtained an accuracy of 92.43 % for the smaller dataset [12]. To classify the defected wheat granules for a dataset of 300 images, an artificial bee colony, performance tuning artificial neural network, and extreme learning machine techniques are used [40]. A deep learning framework for multi-class fruit detection which includes fruits images along with data augmentation based on Faster RCNN was proposed and the performance was evaluated [41]. For identifying pests and plant diseases in video content, a deep learning-based Faster RCNN was investigated along with video based performance metrics [42]. A survey paper of current innovations in image processing methods for automated leaf pest and disease recognition [43]. Adao et al. collected a dataset of cotton field images and implemented a deep residual design and classified the pests. F1-score of 0.98 was achieved by using Resnet 34 model [44]. A metric for accuracy degradation was utilized to analyze machine learning algorithms by enhancing benign samples [24]. The natural statistics model was applied to create saliency maps and identify regions of interest in an insect pest image. Further work was done on the bio-inspired Hierarchical model and X (HMAX) method in the accompanying areas to retrieve invariant features for representing pest appearance [13]. Convolutional neural network-based frameworks, such as attention, feature pyramid, and fine-grained modeling techniques for the IP102 dataset were implemented and obtained an accuracy of 74.00 % [14]. Chen. H. C et al. implemented the AlexNet-modified architecture-based convolutional neural network model on the mobile application in order to identify tomato diseases utilizing leaf images. For a 9-class disease, the Alexnet model had a precision of 80.3% [15]. Pest detection for 10 pest classes using an efficient system for deep learning achieved an average accuracy of 70.5 %. Yolov5-S model was used for the detection of pests and the dataset used was IP102 [16]. A comparison of KNN, SVM, Multilayer Perceptron, Faster R-CNN, and Single Shot Detector classifiers in distinguishing Bemisia Tabacii embryo and Trialeurodes Vaporariorum embryo tomato pest classes was implemented [17]. K. Thenmozhi used three types of the dataset which include NBAIR, Xie1, and Xie2 for insect classification for 40 classes and 24 classes. Pre-trained deep learning techniques like AlexNet, ResNet, and VGGNet were used for insect classification and fine-tuned with pre-trained

models by utilizing transfer learning and obtained an accuracy of 96.75, 97.47, and 95.97% [18].

Wang et al. implemented a Multi-scale convolution capsule network for crop insect pest detection. The advantages of MSCCN are that it is able to extract the multi-scale discriminative features, encode the hierarchical structure of size-variant pests and for pest identification, softmax function was used to determine the probability. They obtained an accuracy of 89.6% for 9 classes of insect species [19]. Nour et al. worked on the AlexNet model to recognize the pests for an IP102 dataset. The model accuracy was fine-tuned by data augmentation to obtain an accuracy of 89.6 % for an eight-class insect pest [20]. Balakrishnan et al. implemented a real-time IOT-based environment to detect pests using a faster RCNN ResNet50 model for object detection framework. The model used 150 test images for each class of insects, 8 classes of the IP102 dataset. The model average accuracy achieved for eight-class insects is around 94.00 % [21]. Kasinathan et al. implemented machine learning techniques such as ANN, SVM, KNN, Naïve Bayes, and the CNN model for pest detection and classification. The model achieved an accuracy of 91.5 % and 93.9 % for nine class and five class pests. The drawback of this model they have used 50 images for each class even though more images of the pests were available in the dataset of IP102 [22].

Mohamed et al. developed a mobile application that uses deep learning to automatically classify pests and for the identification of insect pests, they used a Faster R-CNN model. The model achieved an average accuracy of 98 % for five pests. The drawback of this work, in training the image pests they have used a total of 500 image pests which results in poor approximation, and few test data will result in an optimistic and high variance estimation of prediction accuracy [23]. In order to overcome the above approach, the proposed work was implemented by using a Faster R-CNN for detection and classification of pests for around 1449 pest images for testing of five pest classes, similarly for 10 classes is 2921 images and 15 classes are 4321 pest images of IP102 dataset.

### A. Insect Pests

The proposed work, includes 15 classes of crop insect pests namely aphids, cicadellidae, flax budworm, flea beetle, cabbage butterfly, peachtree borer, prodenia litura, thrips, bird cherry-oat aphid, mole cricket, grub, wireworm, ampelophaga, lycoma delicatula and xylotrachus. Each class in the IP102 dataset is highly unbalanced, each class pest that contains more images in the dataset is taken into consideration for the study. The following insect pests framework cause considerable damage to the crops leading to a loss in crop productivity.

### B. Faster R-CNN

Faster R-CNN requires an image to be scaled to a certain length and width so that noise can be avoided and with the introduction of a region proposal network the detection speed of insect pests is vastly improved [37]. The feature map is generated by the convolution neural network layers for processing the images and the identified object undergoes location regression and classification. We evaluate the three important steps which are involved in Faster R-CNN. Feature

maps were obtained from a pre-trained convolutional neural networks framework in particular Efficient Net [25], Resnet 50 [21], and Dense convolutional neural networks [38]. Next, the Region Proposal network generates the region proposals to detect the pest's locations in the image. The regression box provides the exact location of the insect pests. The insect pest image processed by the region generative proposal is sent to the region of interest pooling to identify and predict the accurate location of the insect pest image. Fig. 3 depicts the proposed Faster R-CNN framework for detection and classification.

### III. MATERIALS AND METHODS

Efficient Net is a unique scaling method that uniformly scales all depth/width/resolution dimensions using a compound coefficient. Neural architecture search is used to generate a brand-new baseline network and scale it up to create the Efficient Nets family of modeling techniques, which outperform prior convolutional networks in both efficiency and accuracy, reducing parameter size and FLOPS [39]. Width scaling is the process of changing the width of an input image. The larger the image, the more feature maps/channels are possible, and thus the more information is available to process [36]. Resolution scaling is the process of changing the resolution of an image. The higher an image's Dots per inch, the higher its resolution. Better resolution is simply an augmentation in the number of pixels in an image. To scale the three dimensions, a baseline model called Efficient Net B0 was introduced. There are seven Efficient Net models ranging from B0-B7, where B0 is the baseline model. The size of the incoming image varies between models. As the model level increases, so does the image's input size. This flexible scaling strategy can be utilized to effectively scale Convolutional Neural Networks and enhance the accuracy with a variety of frameworks.

The input image is processed by MBConv bottlenecks in which direct connections are used because between bottlenecks with significantly fewer channels than expansion layers in inverted residual blocks as shown in Fig. 1. MB Conv has an attention blocks and are made up of a layer that expands and then compresses the channels, mechanism that allows it to optimize channel features that contain the highest information while restricting less significant channel features. The gradient of MBConv does not quickly vanish when the network depth is more thereby improving the model performance. The regularization effect can be increased by using a swish function with no upper limit wherein gradient saturation will not occur [25]. In order to improve the performance, the network drop connect is used to prevent over-fitting. The Efficient Net B4 and Efficient Net B7 model consists of nine phases with respect to Blocks. Blocks provide effective layers and their feature map is connected to the Region Proposal network and Region of Interest pooling as shown in Fig. 3.

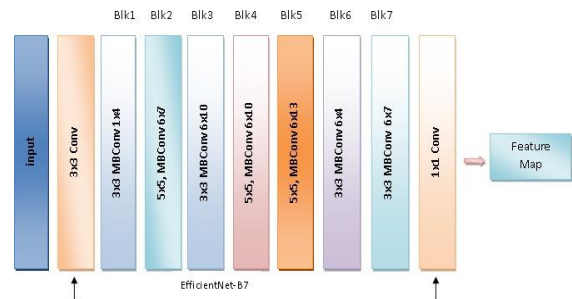


Fig. 1. Efficient B7 Architecture.

#### A. Dataset

Capturing pest images is a difficult task although all insect pests go through several phases during their entire life, based on the species and category of pest. IP102 dataset is commonly used to test the insect pest for classification and detection based on deep learning methods [22]. As a result, we utilized pest images from the public IP102 dataset. The dataset has around 75000 images pertaining to 102 insect pest species. For detection and classification, we have chosen 5, 10, and 15 classes of insect pests. Dataset of 14490 pest images for the training of five pest classes, 29210 images for 10 pest classes, and 43210 images for 15 pest classes. The pest images were split in the ratio of 80 % training, 10 % validation, and 10 % for testing. Sample images of insect pests are shown in Fig. 2.

#### B. Proposed Framework for Detection and Classification

The pest images of the IP102 dataset are passed to Efficient Net network and are pre-trained on ImageNet to generate feature map. In order to improve the performance, network Drop connect and Swish function is utilised in Efficient Net. The feature map is passed to the RPN network to generate the bounding box and proposal score for the pest images. The output of RPN network and feature map obtained from the Efficient Net algorithm is passed to ROI pooling for detection and classification of pest images. Further the flow of the Proposed Faster R-CNN framework for Pest detection and Classification is explained in detail in the below following Section 3C and 3D.



Fig. 2. Samples of Pests Images from the IP102 Dataset.



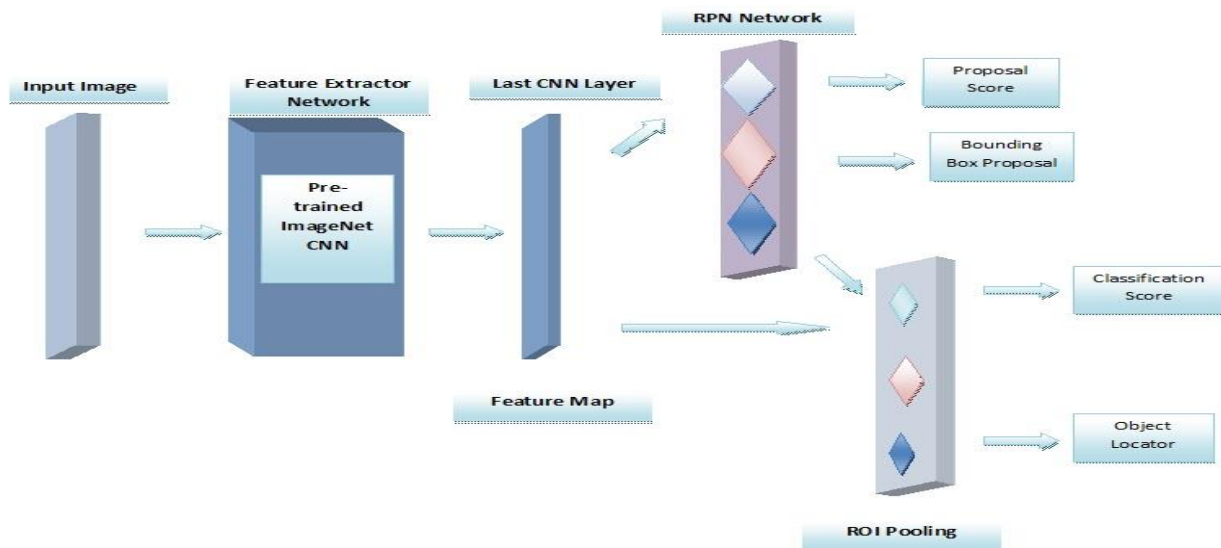


Fig. 3. Proposed Faster R-CNN Framework for Pest Detection and Classification.

### C. Image Preprocessing and Augmentation

Images are transformed to (600,600) in the pre-processing stage phase to retain the same aspect ratio, and images are normalized to maintain the standardized data distribution [25]. The importance of data augmentation for image classification analysis has previously been proven due to insufficient datasets. The categories of each insect pest in the IP102 dataset are highly unbalanced. To increase the data while avoiding the over-fitting problem, various data augmentation techniques such as rescaling, zooming, and horizontal flipping have been used. Gaussian filter is first used to smooth the image. The images were rescaled, created a mask for every image, and then applied segmentation to each sample. Each image in the dataset is subjected to the processing pipelining by a function.

### D. Insect Pest Detection & Classification

The above-proposed learning architecture is used for image processing and to detect and classify pests using the Efficient Net and Faster R-CNN approach as shown in Fig. 3. The convolutional neural network layers of Efficient Net B4 and Efficient Net B7 has been used as feature extractor in this research and for Faster R-CNN because of their added advantage of lightweight and its processing speed which is critical for our end application. The pre-trained weights of Efficient Net were trained on the Image Net dataset. The size of the input image for this methodology is fixed at 224 x 224. Hence using the EfficientNet model we generate feature maps for an input image and pass it to the RPN.

The RPN takes these feature maps as an input to it and provides a set of rectangular proposals (bounding box) identifying the object i.e, a pest in the convolutional neural network feature map as an output along with the objectness score. Grid-anchor having aspect ratio [0.25, 0.5, 1.0, 2.0] is started with a 16x16 pixel size during this stage. These anchors point to available objects of different sizes and aspect ratios at the corresponding location. Intersection over Union determines how well the bounding box matches with the ground truth of the insect pest image, where A and B are two

sections of region proposals as given in (1). To improve the performance of the model and to reduce the noise, non-maximum suppression is utilized for identifying the bounding boxes with the highest confidence so that the small overlaps are ignored. The thresholds were kept at 0.7.

$$IoU = \frac{A \cap B}{A \cup B} \quad (1)$$

To create the proposals for the object, the Faster R-CNN architecture is utilized. It has a specialized and unique architecture that has got classifier and regressor. The Faster R-CNN is robust against translations, it's one of the important properties that it is translational invariant.

When multi-scale anchors are present, Faster R-CNN creates a "Pyramid of Anchors" rather than a "Pyramid of Filters," which consumes less time and is more cost efficient compared to various other architecture. The next step is to pass the proposals to Region of Interest pooling layers. To create a single feature map for each of the proposals provided by RPN in a single pass, Region of Interest pooling is utilized. It is implemented to address the issue of fixed image size difficulties with object detection. ROI pooling is utilized to create fixed-size feature maps against non-uniform inputs by applying max-pooling across the inputs. This layer needs two inputs: (i) A feature map obtained from a backbone of EfficientNet B4 or EfficientNet B7 used in our research methodology after multiple convolutions and pooling layers. (ii) 'N' proposals or Region of Interests from region proposal network (RPN).

The benefit of Region of Interest pooling is that we can utilize the corresponding feature map across all proposals, allowing us to pass the whole image to the convolutional neural networks rather than passing each proposal separately. The sub-windows have a size of (N, 7, 7, 512) which has been created by the Region of Interest pooling layer by applying max pooling over the next stage, where N represents the number of region proposals obtained by the RPN network. The features are moved into the classifier and regression sections after moving via two fully connected

layers. Using the softmax function, the classification division evaluates the probability of a region proposal comprising an insect pest. Additionally, Intersection over Union values are used to evaluate the accuracy of the bounding box generated surrounding the insect pest. The anchor box coordinates are provided by the bounding box regression.

*E. Classification Performance Metrics*

The performance for identifying the insects is measured by using rotation estimation for validating the insect pests of tested images with predicted classification results of the Faster R-CNN technique [26]. The Confusion Metrics are evaluated by True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). The TP indicates the current predicted insect pest class category that is correctly classified. The TN pertains to other groups that do not belong to the existing insect pest class category. The FP pertains to other insect pest class category incorrectly classified as the current insect pest class type. The FN relates to the current insect pest class category that was incorrectly classified and did not belong to the existing class. Precision metric indicates out of all points that are predicted to be positive, how many are actually Positive. The recall metric indicates out of all positive points, how many are actually positive. The classification metrics is given below.

$$accuracy = \frac{tp+tn}{tp+fp+tn+fn} \tag{2}$$

$$precision = \frac{tp}{tp+fp} \tag{3}$$

$$recall = \frac{tp}{tp+fn} \tag{4}$$

$$f1score = \frac{2*precision*recall}{precision+recall} \tag{5}$$

IV. RESULTS AND DISCUSSIONS

For this experiment, we have used an i7 processor with GPU (Nvidia RTX 3080 Ti) along with other supporting tools such as Keras and Tensor flow for the detection and classification analysis of insect pest images of the IP102 dataset. The performance of the insect detection and classification method was implemented on 5, 10 and 15 classes of insects. The insect pest images were split into the ratio of 80% training, 10% validation, and 10 % for testing. The proposed Faster R-CNN model is trained using Stochastic Gradient Descent as an optimizer with 0.9 momentum value, region proposal network weights, and the last fully connected layer weights. The learning rate tells about the learning progress and updates with weight parameters to reduce the loss. The learning rate is varied from 0.0005, 0.0001, 0.001. The maximum no of epochs trained to 40 steps. The detection and classification results are shown in Fig. 4 based on Faster RCNN. The proposed Faster R-CNN technique can correctly detect insect pests in the image and identify the categories. For all test datasets of pest species, classification accuracy ranged from 97.0 to 100.00%.



Fig. 4. Sample of Pest Detection Results for the IP102 Dataset.

The performance indicator for Pest detection is shown in Fig. 5, such as the two methods' average inferential speed. As shown in Fig. 5, Faster R-CNN Efficient Net B7 speed along with accuracy takes around 19.5 frames per second compared to the other model which takes 20.7 frames per second.

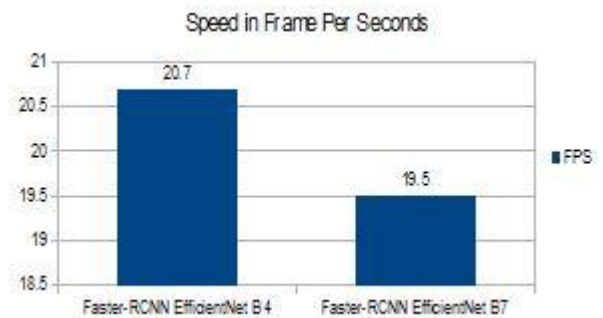


Fig. 5. Speed for Insect Pest Detection based on Faster-RCNN.

The model performance for five pest classes based on Faster R-CNN Efficient Net B7 and Faster R-CNN Efficient Net B4 model is shown in Fig. 6 and Fig. 7. The learning rate was reduced by a factor of 0.5 when the improvement during training went negative. The model continued to be trained with a stop patience of seven, i.e, if for seven continuous epochs there was a negative improvement, the training was halted automatically. The Validation accuracy of around 99.00 % was achieved during training, and the validation loss decreased progressively up to 0.4 % for the Faster R-CNN Efficient Net B7 model. Similarly, the Validation accuracy of 98.00 % and loss of 0.6 % are obtained for the Faster R-CNN Efficient Net B4 model.

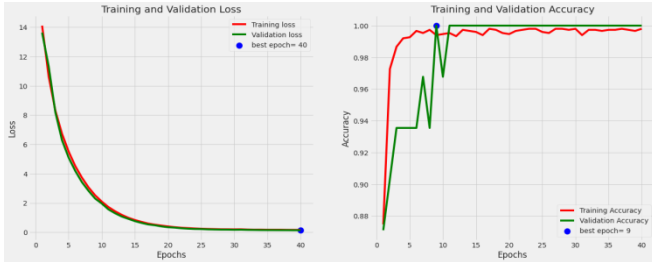


Fig. 6. Model Performance for 5 Pest Classes based on Faster R-CNN Efficient Net B7.

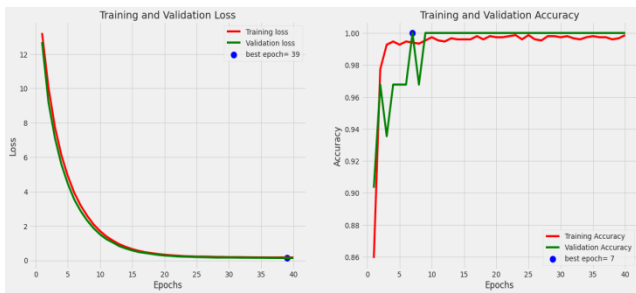


Fig. 7. Model Performance for 5 Pest Classes based on Faster R-CNN Efficient Net B4.

The model performance for 10 pest classes based on Faster R-CNN Efficient Net B7 and Faster R-CNN Efficient Net B4 model is shown in Fig. 8 and Fig. 9. The Validation accuracy of around 96.00 % was achieved during training and the validation loss decreased progressively up to 0.6 % for the Faster R-CNN Efficient Net B7 model. Similarly, the validation accuracy of 95.00 % and loss of 0.7 % are obtained for the Faster R-CNN Efficient Net B4 model.

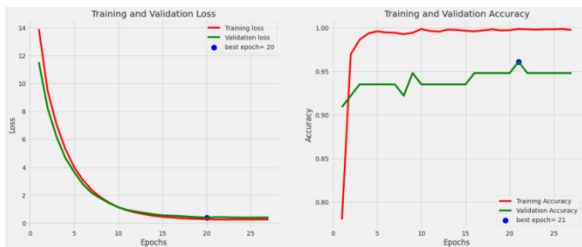


Fig. 8. Model Performance for 10 Pest Classes based on Faster R-CNN Efficient Net B7.



Fig. 9. Model Performance for 10 Pest Classes based on Faster R-CNN Efficient Net B4.

Similarly, we investigated the model performance for 15 pest classes based on Faster R-CNN Efficient Net B7 and the Faster R-CNN Efficient Net B4 models as shown in Fig. 10 and Fig. 11. The Validation accuracy of around 93.00 % was achieved during training and the validation loss decreased progressively up to 0.67 % for the Faster R-CNN Efficient Net B7 model. The validation accuracy of 86.00 % and loss of 0.72 % are obtained for the Faster RCNN Efficient Net B4 model.

Fig. 12, Fig. 13, and Fig. 14 shows the confusion matrix for 5, 10, and 15 Pest classes of the IP102 dataset during testing for the Faster R-CNN Efficient Net B7 model. For insect pests, aphids 0.006 %, cabbage butterfly 0.019 %, cicadellidae 0.006 % and flea beetle 0.042 % of images is incorrectly classified for five pest class. Flax budworm is correctly classified with a ratio of one for a five pest insect classification.

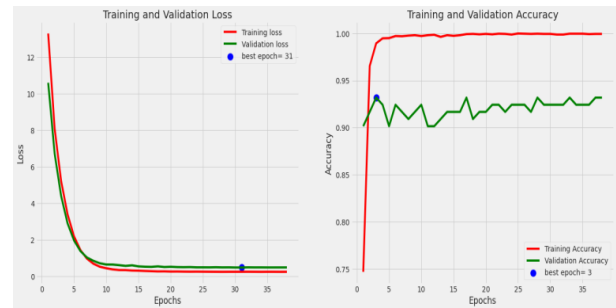


Fig. 10. Model Performance for 15 Pest Classes based on Faster R-CNN Efficient Net B7.

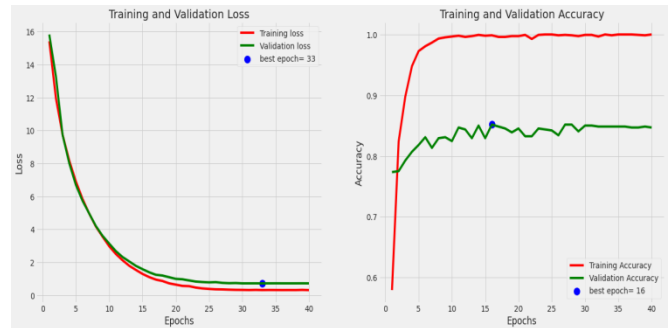


Fig. 11. Model Performance for 15 Pest Classes based on Faster R-CNN Efficient Net B4.



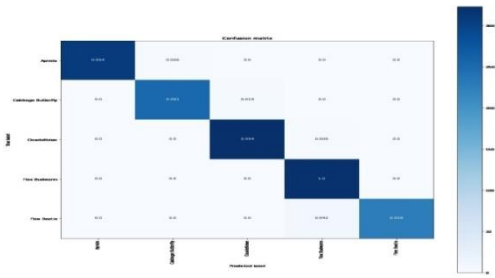


Fig. 12. Confusion Matrix for 5 Pest Classes for Faster R-CNN Efficient Net B7 Model.

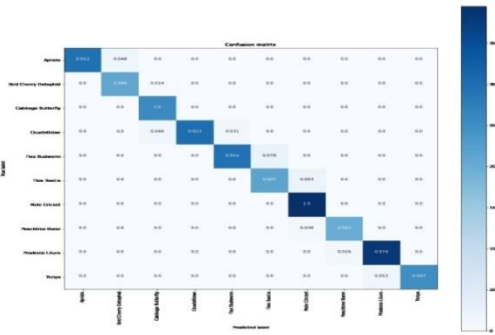


Fig. 13. Confusion Matrix for 10 Pest Classes for Faster R-CNN Efficient Net B7 Model.

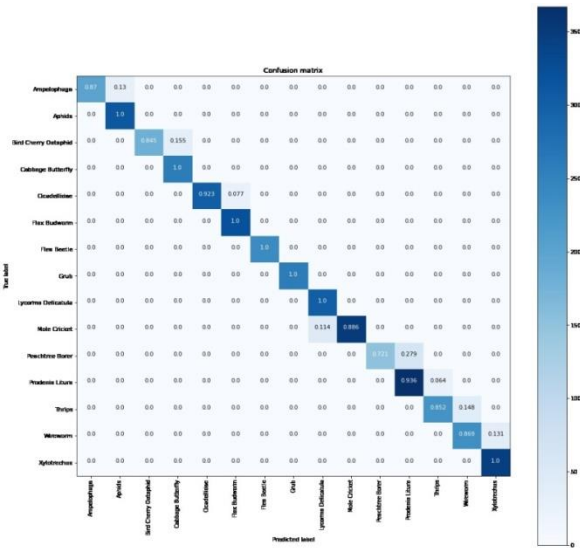


Fig. 14. Confusion Matrix for 15 Pest Classes for Faster R-CNN Efficient Net B7 Model.

Fig. 15 to 17, show the confusion matrix for 5, 10, and 15 Pest classes during testing for the Faster R-CNN Efficient Net B4 model. For insect pest aphids 0.016 %, cabbage butterfly 0.027 %, cicadellidae 0.006 % and flea beetle 0.084 % of images is incorrectly classified for a five pest class. Flax budworm is correctly classified with a ratio of 1 for a 5 pest insect classification.

Fig. 18, illustrates the classification report for the test dataset for 5, 10 and 15 pest classes using Faster R-CNN Efficient Net B7 for IP102 dataset. Classification Accuracy of 99.00 %, 96.00 %, and 93.00 % is achieved for 5, 10, and 15

pest classes based on Faster R-CNN Efficient Net B7. For the five pest classes, the accuracy ranges from 98 % to 100 %. The Precision, recall, the F1 score is 99.00 % for 5 classes and of 10 class pests test data it is around 96.00 % and for 15 class pests test data is 93.00 %.

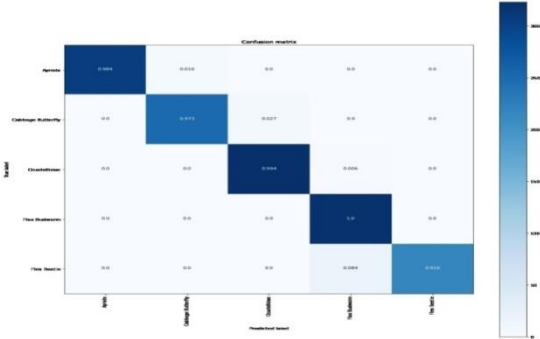


Fig. 15. Confusion Matrix for 5 Pest Classes for Faster R-CNN Efficient Net B4 Model.

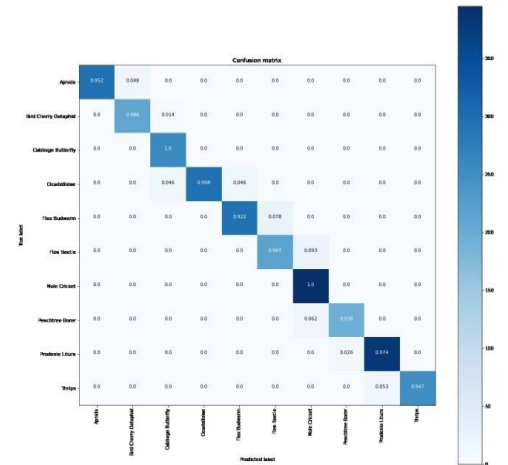


Fig. 16. Confusion Matrix for 10 Pest Classes for Faster R-CNN Efficient Net B4 Model.

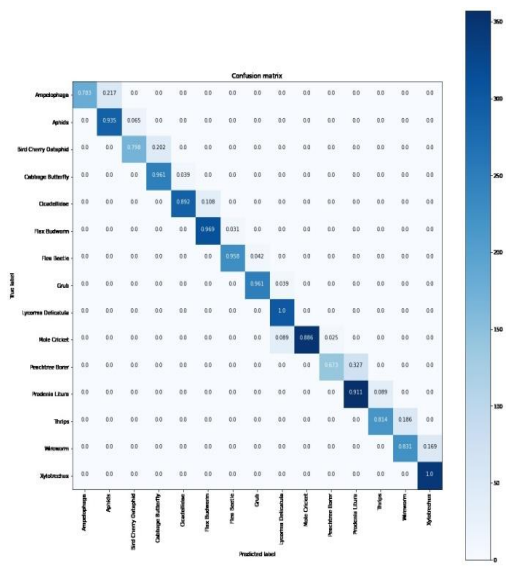


Fig. 17. Confusion Matrix for 15 Pest Classes for Faster R-CNN Efficient Net B4 Model.

	precision	recall	f1-score	support		precision	recall	f1-score	support		precision	recall	f1-score	support
Aphids	1.00	0.99	1.00	310	Aphids	1.00	0.95	0.98	310	Ampelophaga	1.00	0.87	0.93	230
Cabbage Butterfly	0.99	0.98	0.99	257	Bird Cherry Oataphid	0.93	0.99	0.96	213	Aphids	0.91	1.00	0.95	310
Cicadellidae	0.98	0.99	0.99	325	Cabbage Butterfly	0.93	1.00	0.97	257	Bird Cherry Oataphid	1.00	0.85	0.92	213
Flax Budworm	0.96	1.00	0.98	320	Cicadellidae	1.00	0.92	0.96	325	Cabbage Butterfly	0.89	1.00	0.94	257
Flea Beetle	1.00	0.96	0.98	237	Flax Budworm	0.97	0.92	0.94	320	Cicadellidae	1.00	0.92	0.96	325
					Flea Beetle	0.90	0.91	0.90	237	Flax Budworm	0.93	1.00	0.96	320
					Mole Cricket	0.93	1.00	0.96	395	Flea Beetle	1.00	1.00	1.00	237
					Peachtree Borer	0.95	0.96	0.96	208	Grub	1.00	1.00	1.00	258
					Prodenia Litura	0.96	0.97	0.97	392	Lycorma Delicatula	0.87	1.00	0.93	300
					Thrips	1.00	0.95	0.97	264	Mole Cricket	1.00	0.89	0.94	395
accuracy			0.99	1449						Peachtree borer	1.00	0.72	0.84	208
macro avg	0.99	0.99	0.99	1449	accuracy			0.96	2921	Prodenia Litura	0.86	0.94	0.90	392
weighted avg	0.99	0.99	0.99	1449	macro avg	0.96	0.96	0.96	2921	Thrips	0.90	0.85	0.88	264
					weighted avg	0.96	0.96	0.96	2921	Wireworm	0.86	0.87	0.86	267
										Xylotrechus	0.91	1.00	0.95	345
										accuracy			0.93	4321
										macro avg	0.94	0.93	0.93	4321
										weighted avg	0.94	0.93	0.93	4321

Fig. 18. Classification Report for 5, 10 and 15 Pest Classes for Faster R-CNN Efficient Net B7.

Fig. 19, illustrates the classification report for the test dataset for 5, 10 and 15 pest classes using Faster R-CNN Efficient Net B4. Classification Accuracy of 98.00 %, 95.00 %, and 90.00 % is achieved for 5, 10, and 15 pest classes based on Faster R-CNN Efficient Net B4. The Precision, recall, and F1 score is 98.00 % for 5 classes, 10 class pest is around 95.00 % and 15 class pests is 90 %.

#### F. Comparative Analysis

Y. Liu et al. investigated using Back-propagation Neural Network for five pest class dataset of IP102 dataset and achieved an accuracy of 63 %, 50 %, and 43.5 % of classification accuracy for 10 %, 20 %, and 30 % test data set [34]. When compared to BP Neural Network, the Single shot Multi-box detector performed better for identifying the crop pests and achieved an accuracy of 90.6 % for a five pest class [35]. Kasinathan et al. proposed a CNN model for five pest classes and obtained an accuracy of 93.9 % [22]. Our Faster-CNN model outperformed when compared with the other two models for recognizing the pests and obtained a classification accuracy of 99.00 % for 10 % of test data, 98.4 % for 20 % of test data, and 95.5 % for test data. When the training of the

pest images is increased by 70 % to 90 %, the classification accuracy improves. When compared to 30 % of test data our Faster R-CNN model has an accuracy of 95.5%, whereas for BP Neural Network and SSD Mobile Net is around 43.5 % and 85.70 % as shown in Fig. 20.

Comparison was done for the other existing methods for 9 and 10-class crop pests as shown in Fig. 21. Among all these models, bio-inspired methods achieved an accuracy score of 92.50 % [12], with inference we can say that these models used deep learning methodology to detect crop pests. Our proposed Faster R-CNN model outperforms other existing methods and achieved an average accuracy score of 96.00 % for a 10-class crop pest test dataset.

The performance of the proposed method Faster R-CNN Efficient Net B7 and Faster R-CNN Efficient Net B4 is compared with existing methods for the IP102 dataset as shown in Table I. From Table I we can infer that the proposed Faster R-CNN Efficient Net B7 method outperforms the latest competitive approaches in terms of Accuracy for 5 and 10 class crop pests.

	precision	recall	f1-score	support		precision	recall	f1-score	support		precision	recall	f1-score	support
Aphids	1.00	0.98	0.99	310	Aphids	1.00	0.95	0.98	310	Ampelophaga	1.00	0.78	0.88	230
Cabbage Butterfly	0.98	0.97	0.98	257	Bird Cherry Oataphid	0.93	0.99	0.96	213	Aphids	0.85	0.94	0.89	310
Cicadellidae	0.98	0.99	0.99	325	Cabbage Butterfly	0.93	1.00	0.97	257	Bird Cherry Oataphid	0.89	0.80	0.84	213
Flax Budworm	0.94	1.00	0.97	320	Cicadellidae	1.00	0.91	0.95	325	Cabbage Butterfly	0.85	0.96	0.90	257
Flea Beetle	1.00	0.92	0.96	237	Flax Budworm	0.95	0.92	0.94	320	Cicadellidae	0.97	0.89	0.93	325
					Flea Beetle	0.90	0.91	0.90	237	Flax Budworm	0.90	0.97	0.93	320
					Mole Cricket	0.92	1.00	0.96	395	Flea Beetle	0.96	0.96	0.96	237
					Peachtree Borer	0.95	0.94	0.94	208	Grub	0.96	0.96	0.96	258
					Prodenia Litura	0.96	0.97	0.97	392	Lycorma Delicatula	0.87	1.00	0.93	300
					Thrips	1.00	0.95	0.97	264	Mole Cricket	1.00	0.89	0.94	395
accuracy			0.98	1449						Peachtree Borer	0.93	0.67	0.78	208
macro avg	0.98	0.97	0.98	1449	accuracy			0.95	2921	Prodenia Litura	0.84	0.91	0.87	392
weighted avg	0.98	0.98	0.98	1449	macro avg	0.95	0.95	0.95	2921	Thrips	0.86	0.81	0.84	264
					weighted avg	0.96	0.95	0.95	2921	Wireworm	0.82	0.83	0.83	267
										Xylotrechus	0.88	1.00	0.94	345
										accuracy			0.90	4321
										macro avg	0.91	0.89	0.89	4321
										weighted avg	0.91	0.90	0.90	4321

Fig. 19. Classification Report for 5, 10 and 15 Pest Classes for Faster R-CNN Efficient Net B4.

### Comparison of 5 Pest classes

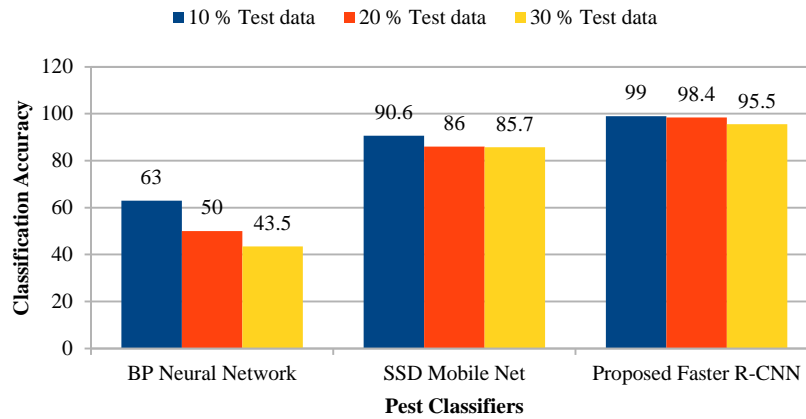


Fig. 20. Comparison of 5 Pest Classes with Existing Methods.

### Comparison for 9 and 10 Pest Classes

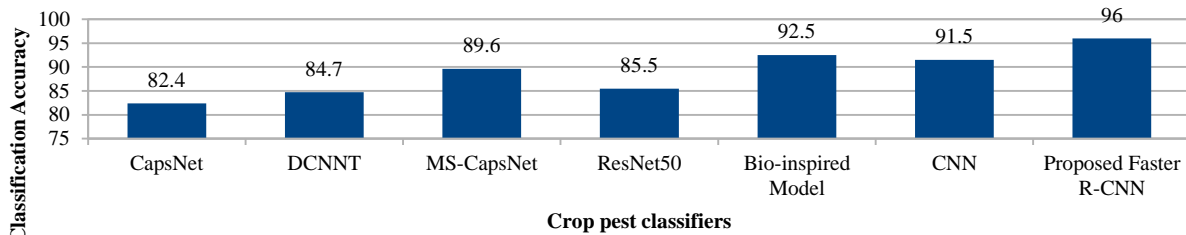


Fig. 21. Comparison of 10 Pest Classes with Existing Methods.

TABLE I. MODEL COMPARISON ON IP102 DATASET FOR CROP PESTS

Research	Technique	Accuracy	Classes
Y. Liu et.al (2016) [34]	BP Neural Network	63.00%	5
W. Liu et.al (2016) [35]	SSD Mobile Net	90.60%	5
Iandola et al. (2016) [27]	SqueezeNet	67.51%	8
Ning et al. (2017) [28]	SSD MobileNet SSD Inception	92.12% 93.47%	8 8
Li et al. (2018) [29]	CapsNet	82.4%	9
Thenmozhi and Reddy (2019) [18]	DCNNT	84.7%	9
Cui et al. (2019) [30]	Yolov2	87.66%	8
Wang et al. (2019) [19]	MS-CapsNet	89.6%	9
Yan et al. (2020) [31]	ResNet50	85.5%	9
Noor et al. (2020) [20]	GoogleNet	88.80%	8
Nanni et al. (2020) [12]	Bio-inspired Model	92.4%	10
Balakrishnan et al. (2020) [21]	Faster-RCNN ResNet50	96.06%	8
Kasinathan et al. (2021) [22]	CNN	91.5% 93.9%	9 5
Chen et al. (2022) [32]	AlexNet	80.3%	9
Xu et al. (2022) [33]	MSCC	92.4%	9
Proposed	Faster RCNN Efficient Net B4 Faster RCNN Efficient Net B7	98.00 %	5
		95.00 %	10
		90.00 %	15
		99.00 %	5
		96.00 %	10
		93.00 %	15



## V. CONCLUSION

In this study, the investigation was done on the Faster R-CNN method to detect and classify different insect pests for 5, 10, and 15 classes, and the results were compared. To improve the performance and accuracy, each one of the pest images has been resized, pre-processed, and augmented to increase the dataset. When the image background is more challenging and the insect classes are more numerous, as in the IP102 dataset, our proposed Faster R-CNN Efficient B7 model achieved an average classification accuracy of 99.00 %, 96.00 %, and 93.00 % for 5, 10, and 15 class insect pests outperforming other existing models such as SSD Mobile Net, Bio-inspired method and Faster R-CNN ResNet 50. In future work, the proposed Faster R-CNN model will be used for higher number of insect classes and subclasses of insect pests that will be useful for farmers to detect insect pests for detection and classification.

## REFERENCES

- [1] Xie, C. Zhang, J. Li, R. Li, J. Hong, P. Xia, J. Chen, P., "Automatic classification for field crop insects via multiple-task sparse representation and Multiple kernel learning "Comput. Electron 2015,119, 123–132.
- [2] Gaston, K. J., "The Magnitude of Global Insect Species Richness," *Conserv. Biol.* 2010, 5, 283–296.
- [3] Siemann E, Tilman. D, Haarstad J, " Insect species diversity, abundance and body size relationships," *Nature* 1996, 380, 704–706.
- [4] Zhang H,Huo Q, Ding W, "The application of AdaBoost-neural network in stored product insect classification," In Proceedings of the IEEE International Symposium on It in Medicine and Education," Xiamen, China, 12–14 December 2009; pp. 973–976.
- [5] Wu X. Zhan C, Lai Y K, Cheng M.M, Yang. J, "IP102: A Large-scale Benchmark Dataset for Insect Pest recognition," In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019, pp. 8779–8788.
- [6] Xie C, Wang R, Zhang J, Chen P, Li R, Chen T,Chen H," Multi-level learning features for automatic classification of field crop pests," *Comput. Electron. Agric.* 2018, 152, 233–241.
- [7] Deng L, "Research on insect pest image detection and recognition based on bio-inspired methods," *Biosystems Engineering. Elsevier*, 169, pp. 139–148.
- [8] Johnny L Miranda,B Gerado, Bartolome T Tanguilg, " Pest Detection and Extraction Using Image Processing Techniques "International Journal of communication and Engineering," DOI:10.7763/IJCCE.2014.V3.317 Corpus ID: 8891485.
- [9] Faithpraise Fina, Philip Birch, Rupert Young , J. Obu , Bassey Faithpraise and Chris Chatwin," Automatic Plant Pest detection and recognition using k-means clustering algorithm and correspondence filters, " *International Journal of Advanced Biotechnology and Research* ISSN 0976-2612, Online ISSN 2278–599X, Vol 4, Issue 2, 2013, pp 189-199.
- [10] Xi cheng, Youhua Zhang, Yigiong Chen, Yunzhi Wu,Yi Yue," Pest identification via deep residual learning in complex background",*Computers and Electronics in agriculture*,volume 141, September 2017, Pages 351-356.
- [11] M.E. Karar, "Robust RBF neural network–based backstepping controller for implantable cardiac pacemakers, " *Int.J. Adapt Control Signal Process.* 32 (2018) 1040–1051.
- [12] Loris Nanni, Gianluca Maguolo, Fabio Pancino," Insect pest image detection and recognition based on bio-inspired methods", *Ecological- Informatics*, <https://doi.org/10.1016/j.ecoinf.2020.101089>. [CrossRef]
- [13] Limiao Deng, Yanjiang Wang, Zhong zhiHan, Renshi Yu, "Research on insect pest image detection and recognition based on bio-inspired-methods," <https://www.sciencedirect.com/journal/biosystems-engineering>.
- [14] Hieu T. Ung , Huy Q. Ung , Binh T.Nguyen."An Efficient Insect Pest Classification Using Multiple, Convolutional Neural Network Based Models", arXiv:2107.12189v1 [cs.CV] 26 Jul 2021.
- [15] ChenH.C, Widodo A.M, Wisnujati A, Rahaman M, Lin J.C.W, Chen L, Weng C.E," AlexNet Convolutional Neural Network for Disease Detection and Classification of Tomato Leaf," *Electronics* 2022, 11, 951. [CrossRef].
- [16] Thanh-Nghi Doan, "An Efficient System for Real-time Mobile Smart Device-based Insect Detection," *International Journal of Advanced Computer Science and Applications*,Vol. 13, No. 6, 2022.
- [17] Gutierrez A, A. Ansuategi, L. Susperregi, C. Tubío, I. Rankić, and L. Lenza. 2019," A benchmarking of learning strategies for pest detection and identification on tomato plants for autonomous scouting robots using internal databases," *Journal of Sensors* 2019.
- [18] Thenmozhi K, Reddy," U.S. Crop Pest Classification Based on Deep Convolutional Neural Network and Transfer Learning," *Comput. Electron. Agric.* 2019, 164, 104906. [CrossRef].
- [19] Wang D, Xu Q, Xiao Y, Tang J, Bin L," Multi-scale Convolutional Capsule Network for Hyperspectral Image Classification," In Chinese Conference on Pattern Recognition and Computer Vision; Springer International Publishing: Cham, Switerland, 2019;pp. 749–760. [CrossRef].
- [20] Nour Khalifa, N.E, Loey M, Taha M,"Insect Pests Recognition Based on Deep Transfer Learning Models," *J. Theor. Appl. Inf. Technol.* 2020, 98, 60–68. [CrossRef].
- [21] Balakrishnan Ramalingam, Mohan R.E, Pookkuttath S, Gómez B.F, Sairam Borusu C.S.C, Wee Teng T, Tamilselvam Y.K," Remote Insects Trap Monitoring System Using Deep Learning Framework and IoT," *Sensors* 2020, 20, 5280. [CrossRef].
- [22] Kasinathan T, Singaraju D, Uyyala S.R, " Insect Classification and Detection in Field Crops Using Modern Machine Learning Techniques," *Inf. Process. Agric.* 2021, 8, 446–457. [CrossRef].
- [23] Mohammed Esmail Karar, Alsunaydi F, Albusaymi S, Alotaibi S," A New Mobile Application of Agricultural Pests Recognition Using Deep Learning in Cloud Computing System,"*Alex. Eng. J.* 2021, 60, 4423–4432. [CrossRef].
- [24] Mamoru Mimura, "Impact of benign sample size on binary classification accuracy", *Expert Systems With applications*,volume211,January 2022,118630
- [25] Tang Yu,Wang Chen, Gao Junfeng and Hua Poxi," Intelligent Detection Method of Forgings Defects Detection Based on Improved EfficientNet and Memetic Algorithm" *IEEE Access*,Digital Object Identifier 10.1109/ACCESS.2022.3193676.
- [26] M. Sokolova, G. Lapalme," A systematic analysis of performance measures for classification tasks," *Inf. Process. Manage.* 45 (2009) 427–437.
- [27] Iandola F.N, Han S.Moskewicz, M.W, Ashraf K, Dally W.J, Keutzer, K," SqueezeNet: AlexNet-level Accuracy with 50x Fewer Parameters and 0.5 MB Model Size," arxiv 2016, arXiv:1602.07360. [CrossRef]
- [28] Ning C, Zhou H, Song Y, Tang J, "Inception Single Shot MultiBox Detector for Object Detection," In Proceedings of the 2017 , IEEE International Conference on Multimedia & Expo Workshops, Hong Kong, China, 10–14 July 2017; pp. 549–554.[CrossRef].
- [29] Li Y, Qian M, Liu P, Cai Q, Li X,Guo J, Yan H, Yu F, Yuan K, Yu J.et ," The Recognition of Rice Images by UAV Based on Capsule Network," *Clust. Comput.* 2018, 22, 9515–9524. [CrossRef].
- [30] Cui J, Zhang J, Sun G, Zheng B," Extraction and Research of Crop Feature Points Based on Computer Vision," *Sensors* 2019, 19,2553. [CrossRef].
- [31] Yan P, Su Y, Tian X," Classification of Mars Lineament and Non-lineament Structure Based on ResNet50," In Proceedings of the 2020 IEEE International Conference on Advances in Electrical Engineering and Computer Applications, Dalian, China, 25–27 August 2020; pp. 437–441.[CrossRef].
- [32] Chen H.C, Widodo A.M, Wisnujati A, Rahaman M, Lin J.C.W, Chen L, Weng C.E,"AlexNet Convolutional Neural Network for Disease Detection and Classification of Tomato Leaf," *Electronics* 2022, 11, 951. [CrossRef].

- [33] Xu C, Yu C, Zhang S, Wang X, "Multi-Scale Convolution-Capsule Network for Crop Insect Pest Recognition," *Electronics* 2022,11, 1630. [CrossRef].
- [34] Y. Liu, W. Jing, L. Xu, "Parallelizing Backpropagation Neural Network Using MapReduce and Cascading model," *Comput. Intell. Neurosci.* 2016 (2016) 2842780. [CrossRef].
- [35] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, A.C. Berg, "SSD: Single Shot MultiBox Detector," in: B. Leibe, J. Matas, N. Sebe, M. Welling (Eds.), *Computer Vision – ECCV 2016*, Springer International Publishing, Cham, 2016, pp. 21–37. [CrossRef].
- [36] A.G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, H. Adam, "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications," *Computer vision and Pattern Recognition arXiv: 1704.4861* (2017).
- [37] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Santiago, Chile, Dec. 2015, pp. 14 40–1448. doi: 10.1109/ICCV.2015.169.
- [38] G. Huang, Z. Liu, L. Van Der Maaten, K.Q. Weinberger, "Densely connected convolutional networks, *Proceedings – 30<sup>th</sup> IEEE Conference on Computer Vision and Pattern Recognition, "CVPR 2017*, Institute of Electrical and Electronics Engineers Inc., 2017, pp. 2261–2269.
- [39] Mingxing Tan, Quoc V. Le, "EfficientNet: Rethinking Model Scaling for convolutional Neural Networks," *Proceedings of the 36 th International Conference on Machine, Learning, Long Beach, California, PMLR 97*, 2019.
- [40] Sabanci, K., "Detection of sunn pest-damaged wheat grains using artificial bee colony optimization-based artificial intelligence techniques", *Journal of the Science of Food and Agriculture*, 100(2), pp.817-824.
- [41] Shaohua Wan, Sotirios Goudos, "Faster R-CNN for multi-class fruit detection using a robotic vision system", *Computer Networks*, vol.168, February 2020.
- [42] Dengshan Li, Rujing Wang, Chengjun Xie, Liu Liu, Jie Zhang, Rui Li, Fangyuan Wang, Man Zhou and Wancai Liu, "A Recognition Method for Rice Plant Diseases and Pests Video Detection Based on Deep Convolutional Neural Network". *Sensors* 2020, 20, 578; doi:10.3390/s20030578.
- [43] Ngugi, L.C, Abewahab, M, Abo-Zahhad, M. "Recent advances in image processing techniques for automated leaf pest and disease recognition "A review. *Inf. Process. Agric.* 2021, 8, 27–51.
- [44] Alves A. N., Souza W. S. R., Borges D. L. "Cotton pests classification in field-based images using deep residual networks", *Computers and Electronics in Agriculture* 2020;174.

# Analysis of Noise Removal Techniques on Retinal Optical Coherence Tomography Images

T.M.Sheeba<sup>1</sup>

Research Scholar

Department of Computer Applications

College of Science and Humanities

SRM Institute of Science and Technology, Kattankulathur,  
Chennai, India

Dr. S. Albert Antony Raj<sup>2</sup>

Associate Professor and Head

Department of Computer Applications

College of Science and Humanities

SRM Institute of Science and Technology, Kattankulathur,  
Chennai, India

**Abstract**—In the biomedical field, automatic disease detection by image processing has become the norm in the current days. For early illness detection, ophthalmologists have explored a variety of invasive and noninvasive procedures. Optical Coherence Tomography (OCT) is a noninvasive imaging technique for obtaining high resolution tomographic images of biological systems. The image quality is degraded by noise, which degrades the performance of noisy image processing algorithms. The OCT images captured with speckle noise and prior to further processing, it is critical to use an effective approach for denoising the image. In this paper, we used Median filter, Average filter or Mean filter, Wiener filter, Gaussian filter and Bilateral filter on OCT images in this paper, and discussed the advantages and drawbacks of each approach. The effectiveness of these filters are compared using the Peak Signal to Noise Ratio (PSNR), Mean Square Error (MSE) and Contrast to Noise Ratio (CNR).

**Keywords**—Average or Mean filter; Bilateral filter; denoising image; Gaussian Filter; Median filter; optical coherence tomography; Wiener filter

## I. INTRODUCTION

Many practitioners have recently embraced Optical Coherence Tomography (OCT) as a means of gathering data from the human eye in order to diagnose problems. In today's world, OCT is a well-known imaging technology that is used to monitor retinal illnesses in the medical industry. Although, during the data acquisition phase of OCT, a grainy prototype known as speckle noise is always present. The attendance of speckle noise in OCT images limits image processing, making patient diagnosis harder for a practitioner. Due to the attendance of speckle noise in OCT images, blood vessels and layer bounds appear to be disconnected. The working approach to obtain OCT images is nearly identical to that for collecting ultrasound images, with the exception of the medium utilised to obtain it. Fig. 1 shows the OCT images of human eye. In the OCT image acquisition technique, light beams are employed as an alternative of the sound beams used in ultrasound imaging [1]. OCT imaging has been shown to play a key responsibility in the diagnosis of disorders associated to the retina and

glaucoma in the medical realm [2-4]. OCT is one of the greatest techniques in the medical sector for finding the inside structure of the retina and high-intention images of the retina [5-8]. It has been discovered via the observations of several specialists that retinal layer width improves experimental results in the field of optometrist. In the case of glaucoma development and macular degeneration, retinal layer segmentation also improves clinical findings. The speckle noise in OCT images degrades picture quality and reduces the image's contrast to noise ratio. A despeckle method is necessary as a pretreatment step in the denoising operation of OCT images to defeat the effect of speckle noise and to maintain the excellent details of the OCT images. Preprocessing OCT images is therefore a crucial step in ophthalmology to improve clinical findings. The motivation of this research work is to choose the best denoising method to reduce speckle noise in the retinal OCT images without removing the details in the image. We analysed the Median Filter, Gaussian Filter, Bilateral Filter, Wiener Filter, and Average or Mean Filter denoising methods on OCT images and proved the Wiener filter is the best denoising method to reduce speckle noise in retinal OCT images.

There are six sections in this paper. Introduction given in Section I, the brief description about the different types of noise is given in Section II, Section III contains denoising methods such as the Median Filter, Gaussian Filter, Bilateral Filter, Wiener Filter, and Average or Mean Filter, the implementation is given in Section IV and Section V discusses the performance measuring methods such as PSNR, CNR, and MSE, as well as how these measurements are used to evaluate results. Section VI concludes with a conclusion.

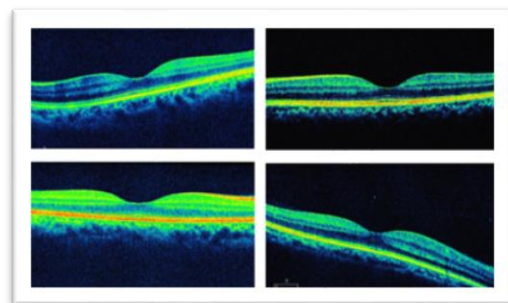


Fig. 1. OCT Images of Human Eye.

## II. LITERATURE REVIEW

OCT is a technique that is frequently used to detect and track retinal conditions. However, despite technology advancements, speckle noise continues to significantly impair its scans. Speckle noise lowers the accuracy of measurements and the dependability of subsequent equipment. Mahnoosh Tajmiriahi et al. used a lightweight convolutional AE (Auto Encoders) used to simulate a recently developed state-of-the-art OCT picture denoising technique [9].

Xiaojun Yu et al. developed a two-step filtering method to reduce speckle noise, which consisting of Augmented Lagrange function minimization and Rayleigh alpha-trimmed filtering (AR) scheme. This mechanism examines the effects of speckle noise distributions on the OCT despeckling process [10].

Nahida Akter et al. proposed deep learning based method to remove noise from OCT images. The authors created and trained a U-Net model using OCT artifact-filled and artifact-free B-scans for investigation and shown that the U-Net performed better in terms of SSIM and PSNR values in removing the artefacts [11].

Bin Qiu et al. developed a deep network architecture. In the study, the authors used a most well-known frequently used modified DnCNN was used to denoise the OCT images [12].

Lirong Zeng et al. suggested progressive feature fusion attention dense network (PFFADN) for removing speckle noise from OCT images. In order to create a residual block, the authors sequentially connected the shallow and deep convolution feature maps that were retrieved from each dense block. This is done by arranging densely connected dense blocks in the deep convolution network [13].

Ling Chen et al., used the SC-based denoising database creation is the central component of the described approach, SC-DnCNN. Since FF-OCT images don't require registration before SC due to their unique image characteristics, they outperform point scanning OCT in terms of producing clear images. This method allows for the inclusion of both noisy and relatively clear images as training data, the embedding of a spatial adaptive mapping based on the compounding database, and the reduction of the effect of the speckle [14].

Yan Hu et al. provided an adaptive-SIN filtering technique to address the problem of minimising the noise in OCT images of various types. The suggested square-root transform converts the Poisson noise in the OCT pictures to the Gaussian noise in order to enable the best noise removal by the subsequent shearlet transform. The edge information in the photos as well as other image fine features may be preserved by the 3D shearlet transform [15].

## III. TYPES OF NOISE

Unwanted information causes image quality to deteriorate. The type of noise there in the original image plays a crucial impact in the image noise removal procedure. Noise having a Gaussian, salt and pepper sharing corrupts typical images. Speckle noise, which is multiplicative in nature, is another example of a typical noise. The following sections detail the behaviour of each of these noises.

### A. Gaussian Noise

Gaussian noise is geometric noise with a normal distribution probability density function, commonly known as Gaussian noise. That is, the noise's possible values are Gaussian-distributed. The noise with a Gaussian amplitude sharing is correctly defined as Gaussian noise [16].

### B. Salt and Pepper Noise

Quick, unexpected perturbations in the image signal generate salt and pepper noise, which appears as at random dotted white or black pixels over the image. In salt and pepper noise, the black pixels appear in bright areas and brightly pixels appear in dark areas. Dead pixels, analogue to digital converter problems, and transmission bit mistakes can all contribute to this form of noise [16].

### C. Speckle Noise

All fundamental aspects of logical imaging, particularly clinical ultra sound imaging, are affected by speckle noise. Sound processing of backscatter signals from several spread targets is the cause. Signals from basic scatters generate speckle noise. Speckle noise is referred to as texture in pharmaceutical literature, and it may include diagnostic information Smoothing the texture may be less desirable for visual interpretation. Physicians prefer the original noisy photos to the smoothed versions more willingly because the filter, even if it is more sophisticated, can eliminate some important image information. As a result, it's critical to create noise filters that maintain the traits that matter to doctors. To reduce speckle noise, several methods are utilised, each based on a distinct mathematical description of the phenomenon. For eliminate speckle noise in ultrasound pictures, we recommend hybrid filtering techniques [16].

## IV. FILTERING TECHNIQUES

A variety of filtering techniques are obtainable in the literature for the elimination of noise. Linear filtering techniques and non-linear filtering techniques are the two main categories. A linear-output filter's is the same as the input, whereas a non-linear-output filter's is different.

### A. Mean Filter or Average Filter

The mean filter counts each pixel in a picture by replacing it with the mean or average value of its neighbours. So in the output image the pixel values that are out of character with their surroundings are deleted. It's built on a kernel, which, like other convolutions, represents the structure and width of the sampled neighbourhood when computing the mean. As shown in Fig. 2, a  $3 \times 3$  kernel is commonly employed, but greater kernels, such as  $5 \times 5$ , may be utilised for more severe smoothing.

### B. Median Filter

Median filtering is a nonlinear smoothing technique. Restore the value of each pixel with the median value of neighborhood element, which is better ordered for decreasing salt and pepper noise. That is, we choose a kernel and sort all elements in the neighbourhood by grey value, with the group's median acting as the output element for the neighbourhood centre. When an element in the proximity of other elements is

unusual, use the grey value of the elements in the field. That is the value that is halfway between the two extremes [17]. Take the field's element grey value when a element neighbourhood of element number is even. In the middle of two scales, it is the sort value. The median filter reduces noise while maintaining image edge clarity. When the window size is increased in median filtering, noise is successfully reduced [18].

$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{9}$
$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{9}$
$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{9}$

Fig. 2. 3x3 Average Kernel Often used in Mean Filter.

### C. Max and Min Filter

In max and min filter, it assigns a new value to each pixel in an image based on the greatest or smallest amount of value in the neighbourhood around that pixel. The filter's shape is represented by the neighbourhood. Contrast enhancement and normalization [19], texture description [20], edge detection [21-22], and thresholding [26] have all employed maximum and minimum filters. The filters are dilation and erosion counterparts with a grey value.

### D. Gaussian Filter

Gaussian filters are thought to be the best time domain filters. It's a type of lowpass filter that isn't uniform. Such filters have a Gaussian impulse response, and Gaussian filters are those that have a Gaussian function. It has the shortest feasible grouping delay. The fundamental purpose of a Gaussian filter is to reduce distortion in lowest and highest signals [23]. A Gaussian obscure, also known as Gaussian smoothing is an after-effect of concealing a picture with a Gaussian capability in image processing. This Gaussian smoothing superintendent is usually a 2-D convolutional superintendent that is used to obfuscate images and remove subtle element and clamours. Gaussian filter is frequently more difficult with salt and pepper. When compared to other filters, one of the key disadvantages of the Gaussian filter is that it takes a long time. As a rule, Gaussian filters do not overshoot a stage work input while restricting the climb and fall times [24].

### E. Bilateral Filter

The bilateral filter is defined as a loaded average of pixels, similar to the Gaussian convolution. The bilateral-filter, on the other hand, maintains edges by taking intensity variations into account. Bilateral filtering is based on the concept that two pixels are adjacent not only if they engage adjacent spatial regions, but also if their photometric ranges are comparable [25].

### F. Wiener Filter

Inverse filtering, sometimes known as generalised inverse-filtering, is a technique for restoring deconvolution. Inverse-filtering or generalised inverse-filtering can be used to improve an image that is dim using a known lowpass filter. Conversely, inverse filtering is extremely intuitive to additive-noise. We

can create a restoration method for each type of deterioration and then join them using the "one degradation at a time" technique. Inverse filtering and noise smoothing are most stable when Wiener filtering is used. It reduces additive noise while also inverting blurring. In terms of MSE, Wiener filtering is the best option. It minimizes the overall mean square error. A linear-estimation of the unique image is used in Wiener filtering. A probabilistic foundation underpins the procedure. The Wiener-filter in the Fourier-domain can be written as follows because of the orthogonal principle:

$$W(f_1, f_2) = \frac{H^*(f_1, f_2)S_{xx}(f_1, f_2)}{|h(f_1, f_2)|^2 S_{xx}(f_1, f_2) + S_{\eta\eta}(f_1, f_2)} \quad (1)$$

where  $S_{xx}(f_1, f_2)$  and  $S_{\eta\eta}(f_1, f_2)$  are correspondingly value spectra of the initial image and the additive- noise, and  $H(f_1, f_2)$  is the blurring- filter. As can be seen, the Wiener filter has two separate parts: a noise smoothing part and inverse-filtering part. It not only uses inverse filtering (highpass filtering) to deconvolve the data, but it also uses compression to reduce the noise (lowpass filtering).

## V. IMPLEMENTATION

The implementation process of comparison is depicted in the block diagram, Fig. 3. On six separate OCT images of an eye, we employ five noise removal techniques: Mean or Average, Median, Bilateral, Gaussian, and Wiener filters, and the efficiency of each filter is evaluated using PSNR, CNR, and MSE values.

The filtering techniques are implemented using Python. The original images are converted into grayscale images and then the filtering techniques are applied on those grayscale images. The sample resultant filtered images are given in Fig. 4.

## VI. EVALUATION AND RESULTS

### A. Mean-Squared-Error (MSE)

The MSE value is the Mean-squared-error between the enhanced image  $Y(I, j)$  and the actual image  $I(i, j)$ , which for a high-quality image is negligible. The MSE of a given image of range  $M \times N$  is calculated as follows:

$$MSE = \frac{1}{MN} \sum_{i=0}^{M-1} \sum_{j=0}^{N-1} [I(i, j) - Y(i, j)]^2 \quad (2)$$

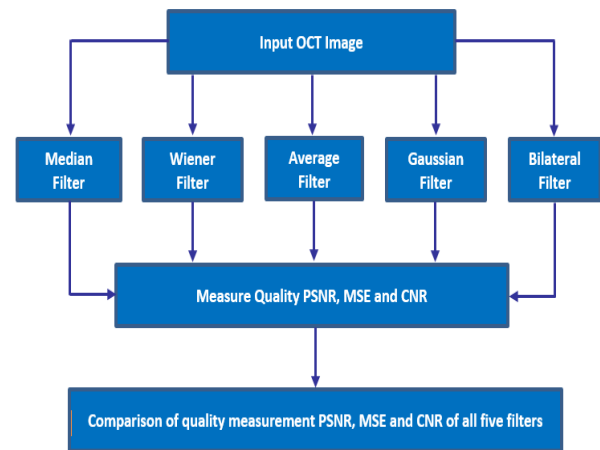


Fig. 3. Block Diagram of Implementation Process of Comparison.

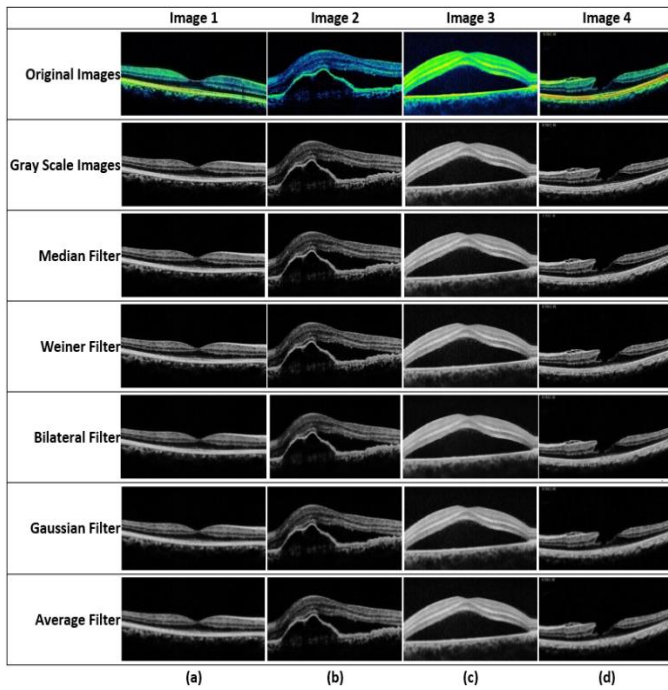


Fig. 4. Four Images are Passing through All Five Filtering Technique before and after, (a) OCT Image 1, (b) OCT Image 2, (c) OCT Image 3, (d) OCT Image 4.

It is determine by squaring the difference between the denoised (images after filtering) and noisy images and then taking the average of the difference [17]. Table I shows the MSE values of various filtered images. Fig. 5 shows the MSE graph after applying the median, wiener, bilateral, Gaussian, and average filters on all six images. Lower values of MSE specify improved image quality. When compared to the other four filters, Wiener has the lowest MSE value.

TABLE I. MSE VALUE OF FILTERED IMAGES

MSE Comparison						
Filters	OCT Image 1	OCT Image 2	OCT Image 3	OCT Image 4	OCT Image 5	OCT Image 6
Median	20.29753	21.86683	30.02484	20.54619	21.85792	27.43176
<b>Weiner</b>	<b>17.51171</b>	<b>19.10760</b>	<b>29.29777</b>	<b>17.57836</b>	<b>18.47427</b>	<b>26.82416</b>
Bilatera l	26.06070	21.61373	39.38961	23.83720	19.67814	35.18031
Gaussia n	27.99892	30.67310	45.20848	29.51530	29.22160	40.63345
Averag e	28.29255	31.12878	46.03533	29.96165	29.39583	41.33654

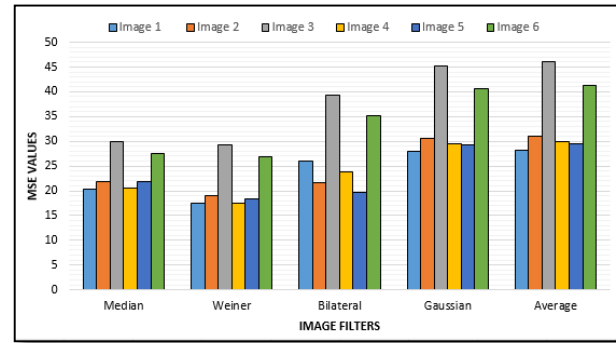


Fig. 5. Comparison of MSE Values of the Filtered Images.

### B. Peak Signal-To-Noise Ratio (PSNR)

PSNR is defined as the relational of a signal's greatest value to the noise value that affects the signal's quality. PSNR is an image quality metric that provides a metric of noise that affects image quality. It determines the quantity of noise that has an impact on image quality. The image quality improves as the PSNR score rises. The PSNR value is calculated using the root-to- mean-square-error (RMSE), which is the square- root of the error between the input-image and the enhanced- image. For the largest possible pixel values, if RMSE will decrease, the PSNR will increase. As a result, greater PSNR values suggest better image improvement. PSNR has the advantage of being computationally and logically straightforward to optimise. One downside of root-mean-squared error is that it changes in lockstep with image intensity. PSNR's mathematical formulation is as follows:

$$PSNR = 10\log(S^2/MSE) \quad (3)$$

The greatest pixel value that can be assign in an image is S. For an 8-bit image, S=255. It can alternatively be defined as the ratio of greatest signal power to maximum noise power. This rate is expressed in decibels (dB). Table II shows the PSNR rate of all six images after going through various filters and Fig. 6 shows the comparison of PSNR values after applying the median, wiener, bilateral, Gaussian, and average filters on all six images. When compared to the other four filters, Wiener has the highest PSNR score.

TABLE II. PSNR VALUE OF FILTERED IMAGES

PSNR Comparison						
Filters	OCT Image 1	OCT Image 2	OCT Image 3	OCT Image 4	OCT Image 5	OCT Image 6
Median	35.0563 dB	34.7329 dB	33.3559 dB	35.0034 dB	34.7347 dB	33.7482 dB
<b>Weiner</b>	<b>35.6975 dB</b>	<b>35.3187 dB</b>	<b>33.3666 dB</b>	<b>35.6810 dB</b>	<b>35.4651 dB</b>	<b>33.8455 dB</b>
Bilateral	33.9709 dB	34.7835 dB	32.1769 dB	34.3582 dB	35.1909 dB	32.6678 dB
Gaussian	33.6593 dB	33.2632 dB	31.5786 dB	33.4303 dB	33.4737 dB	32.0419 dB
Average	33.6140 dB	33.1991 dB	31.4998 dB	33.3651 dB	33.4479 dB	31.9674 dB



C. Contrast to Noise Ratio (CNR)

CNR is a metric for determining the quality of an image. The measure SNR is a like to the CNR, with the exception that it subtracts a term before calculating the ratio [27]. When there is a considerable bias in an image, such as from haze [28], this is critical. The intensity is quite strong, as can be seen in the image to the right, even though the image's features are washed out by the haze. As a result, while this image has a higher SNR measure, it has a lower CNR measure. One method to define difference to noise rate is [29, 30]:

$$C = |S_A - S_B| / \sigma_o \tag{4}$$

$S_A$  and  $S_B$  are signal strengths for signal produce structures A and B in the region of interest, while  $\sigma_o$  is the SD of the clean image noise. Table III shows the contrast to noise ratio of the six images after processing through various filters and Fig. 7 shows the CNR graph after applying the median, wiener, bilateral, Gaussian, and average filters on all six images. Lower values of CNR specify improved image quality. CNR value for Wiener is least as compare to other four filters.

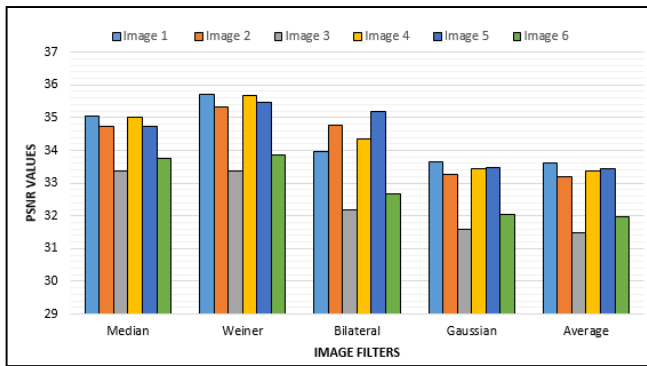


Fig. 6. Comparison of PSNR Values of the Filtered Images.

TABLE III. CNR VALUE OF FILTERED IMAGES

CNR Comparison						
Filters	OCT Image 1	OCT Image 2	OCT Image 3	OCT Image 4	OCT Image 5	OCT Image 6
Median	0.01458	0.02405	0.01017	0.01696	0.01680	0.01045
Weiner	<b>0.00081</b>	<b>0.00166</b>	<b>0.00052</b>	<b>0.00103</b>	<b>0.00029</b>	<b>0.00095</b>
Bilateral	0.01779	0.00892	0.00303	0.00484	0.00475	0.00244
Gaussian	0.00334	0.00436	0.00198	0.00359	0.00323	0.00217
Average	0.00345	0.00441	0.00203	0.00365	0.00332	0.00221

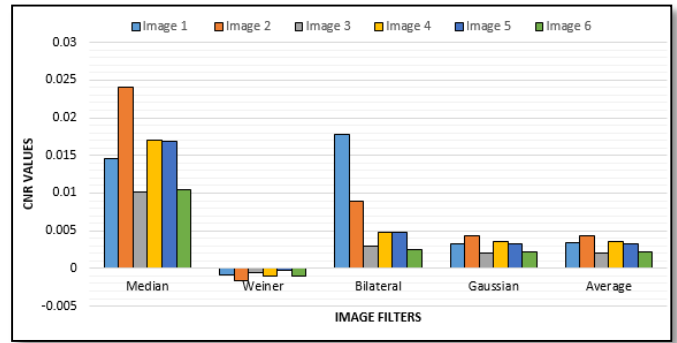


Fig. 7. Comparison of CNR Values of the Filtered Images.

VII. CONCLUSION

Due to the chaos in OCT images, ophthalmologists have difficulty in correctly detecting disease. It is also a barrier to the automatic segmentation of biomedical images for illness diagnosis. The Wiener filter functioned well on all six OCT images, according to the results in Section V. The Wiener filtering algorithm significantly decreases speckle noise, while also preserving retinal formation and reducing the stairway effect. Tables I, II, and III show that among the several despeckling filters evaluated here, the Wiener filtering algorithm is the most excellent way for reducing the produce of speckle noise while keeping the edges. Further, this work can be extended by analysing other denoising methods and quality matrices such as SSIM, SNR, ENL and MAE on retinal OCT images.

REFERENCES

- [1] D. Huang, E. A. Swanson, C. P. Lin, J. S. Schuman, W. G. Stinson, W. Chang "Optical Coherence Tomography," Science, vol. 254, pp. 1178-1181, 1991.
- [2] M. R. Hee, C. A. Puliafito, C. Wong, J. S. Duker, E. Reichel, B. Rutledge "Quantitative assessment of macular edema with optical coherence tomography," Archives of ophthalmology, vol. 113, pp. 1019-1029, 1995.
- [3] R. Koproński and Z. Wróbel, "Image processing in optical coherence tomography using Matlab", Poland: Katowice, 2011.
- [4] A. G. Podoleanu, "Optical Coherence Tomography" The British Journal of Radiology, vol. 78, pp. 976-988, 2005.
- [5] W. Drexler, J.G. Fujimoto, "Optical Coherence Tomography", Springer, 2008.
- [6] Y. Chen, L.N. Vuong, J. Liu, J. Ho, V.J. Srinivasan, I. Gorczynska, A.J. Witkin, J.S. Duker, J. Schuman, J.G. Fujimoto "Three-dimensional ultra high resolution optical coherence tomography imaging of age-related macular degeneration", Opt. Express 17, pp. 4046-4060, 2009.
- [7] I. Krebs, S. Hagen, W. Brannath, P. Haas, I. Womastek, G. deSalvo, S. Ansari-Shahrezaei, S. Binder "Repeatability and reproducibility of retinal thickness measurements by optical coherence tomography in age-related macular degeneration", Ophthalmology, 117, 1577-1584, 2010.
- [8] M. Anand and Dr. C. Jayakumari, "Study of retina image segmentation algorithms from optical coherence tomography(OCT) images", Jour of Adv Research in Dynamical & Control Systems, Vol. 9, No. 4, pp. 125-134, 2017.

- [9] M. Tajmirriahi, R. Kafieh, Z. Amini and H. Rabbani, "A lightweight mimic convolutional auto-encoder for denoising retinal optical coherence tomography images", in *IEEE Transactions on Instrumentation and Measurement*, Vol. 70, pp. 1-8, 2021.
- [10] X. Yu, C. Ge, Z. Fu, M. Z. Aziz and L. Liu, "A two-step filtering mechanism for speckle noise reduction in OCT images," 2021 IEEE 9th International Conference on Information, Communication and Networks (ICICN), pp. 501-505, 2021.
- [11] N. Akter, S. Perry, J. Fletcher, M. Simunovic and M. Roy, "Automated Artifacts and Noise Removal from Optical Coherence Tomography Images Using Deep Learning Technique," 2020 IEEE Symposium Series on Computational Intelligence (SSCI), pp. 2536-2542, 2020.
- [12] Bin Qiu, Zhiyu Huang, Xi Liu, Xiangxi Meng, Yunfei You, Gangjun Liu, Kun Yang, Andreas Maier, Qiusi Ren and Yanye Lu, "Noise reduction in optical coherence tomography images using a deep neural network with perceptually-sensitive loss function", *Biomedical Optics Express*, Vol. 11, No. 2, pp. 817-830, 2020.
- [13] L. Zeng, M. Huang, Y. Li, Q. Chen and H. -N. Dai, "Progressive Feature Fusion Attention Dense Network for Speckle Noise Removal in OCT Images," in *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2022.
- [14] I. -L. Chen, T. -S. Ho and C. -W. Lu, "Full Field Optical Coherence Tomography Image Denoising Using Deep Learning with Spatial Compounding," 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI), pp. 1975-1978, 2020.
- [15] Yan Hu, Jianfeng Ren, Jianlong Yang, Ruibing Bai & Jiang Liu, "Noise reduction by adaptive SIN filtering for retinal OCT images", *Scientific Reports* 11, 19498, 2021.
- [16] Sheikh Tania and Raghada Rowaida, "A Comparative Study of Various Image Filtering Techniques for Removing Various Noisy Pixels in Aerial Image", *International Journal of Signal Processing, Image Processing and Pattern Recognition* Vol.9, No.3, pp.113-124, 2016.
- [17] R. Ramani, "The Pre-Processing Techniques for Breast Cancer Detection in Mammography Images", *I.J. Image, Graphics and Signal Processing*, Vol. 5, pp. 47-54, 2013.
- [18] Rakesh M.R1, Ajeya B2, Mohan A., "Hybrid Median Filter for Impulse Noise Removal of an Image in Image Restoration", *Research in Science*, Vol. 3, Issue 3, 2014.
- [19] Dorst, L., "Quantitative Analysis of Interferograms Using Image Processing Techniques", *ICO-13 Conf. Digest*, Sapporo, Japan 1984, pp. 476-477.
- [20] Werman, M. and S. Peleg, "Min-Max Filters in Texture Analysis", *IEEE PAMI-7*, pp. 730- 733, 1986.
- [21] Lee, J.S.J., R.M. Haralick and L.G. Shapiro, "Morphologic Edge Detection", *Proc. 8th ICPR*, Paris 1986, pp. 369-373.
- [22] Van Vliet, L.J., I.T. Young and A.L.D. Beckers, "A Nonlinear Laplace Operator as Edge Detector in Noisy Images", submitted to *CVGIP*, 1987.
- [23] Nithya. K, Aruna. A, Anandakumar. H, Anuradha. B, "A Survey On Image Denoising Methodology On Mammogram Images", *International Journal of Scientific & Technology Research*, Vol. 3, Issue 11, pp. 92-93, 2014.
- [24] Kshema, M. J. George and D. A. S. Dhas, "Preprocessing filters for mammogram images: A review," 2017 Conference on Emerging Devices and Smart Systems (ICEDSS), pp. 1-7, 2017.
- [25] Sylvain Paris, Pierre Kornprobst, Jack Tumblin, Fredo Durand, "A Gentle Introduction to Bilateral Filtering and its Applications", *ACM Digital Library*, 2007.
- [26] Bensen, J., "Dynamic Thresholding of Grey-Level Images", *Proc. 8th ICPR*, Paris 1986, pp. 1251- 1255.
- [27] Welvaert, Marijke; Rosseel, Yves; Yacoub, Essa. "On the Definition of Signal-To-Noise Ratio and Contrast-To-Noise Ratio for fMRI Data", *PLOS ONE*. 8 (11): e77089 Bibcode: 2013.
- [28] Jiang, Hou; Lu, Ning; Yao, Ling, "A High-Fidelity Haze Removal Method Based on HOT for Visible Remote Sensing Images". *Remote Sensing*. 8 (10): 844, pp. 1-18, 2016.
- [29] Desai, Nikunj, "Practical Evaluation of Image Quality in Computed Radiographic (CR) Imaging Systems", *Proceedings of SPIE - The International Society for Optical Engineering*, 7622, 2010.
- [30] Timischl F. The contrast-to-noise ratio for image quality evaluation in scanning electron microscopy, *Scanning*. 37 (1): 54-62, 2015.

# Analyzing the State of Mind of Self-quarantined People during COVID-19 Pandemic Lockdown Period: A Multiple Correspondence Analysis Approach

Gauri Vaidya<sup>1</sup>, Vidya Kumbhar<sup>2</sup>, Sachin Naik<sup>3</sup>, Vijayatai Hukare<sup>4</sup>

Symbiosis Institute of Geoinformatics, Symbiosis International (Deemed University), Pune, Maharashtra, India<sup>1, 2, 4</sup>

Symbiosis Institute of Computer Studies & Research, Symbiosis International (Deemed University), Pune, Maharashtra, India<sup>3</sup>

**Abstract**—COVID-19 (Corona) virus has spread across the world threatening lives of millions of people. In India first COVID-19 case was detected on 30th January 2020 in Kerala. To minimize the spread of Corona Virus, countries all over the world implemented complete lockdown. Due to complete lockdown even people who are not exposed to corona virus, have to self-quarantine to keep themselves safe from getting infected by the disease. People (especially Indians) have never experienced such complete lockdown and quarantining situations before. Thus, this creates a space for curiosity that how people are going to react to this situation. The present study aims to analyse how self-quarantined people during COVID-19 lockdown period are reacting to quarantining, what measures they are taking to deal with this situation, and what are their sentiments towards quarantining. The study also aims to measure their Happiness and to identify the factors that are statistically significant to Happiness. For this study, the data is collected through a survey method. Multiple correspondence analysis are performed to analyse the survey data. The happiness score is evaluated by using the GNH (Gross National Happiness) methodology. Proportional Odd Logistics Regression is used to identify factors that are statistically significant in predicting happiness. The study suggests that family factor is related to the happiness of the self-quarantined people during such lockdown situations.

**Keywords**—Correspondence analysis; happiness index; sentiment analysis; proportional odds logistic regression; self-quarantining

## I. INTRODUCTION

‘Quarantine’ is a practice in which restrictions are imposed on the movement of people. Sometimes people may have been exposed to disease but do not show any symptoms of being infected by the disease. In such cases, people are not allowed to go to public places for some period. This period is called as quarantine period. Quarantine period identifies whether someone is being infected by the disease or not. In case of COVID-19, the Corona Virus has spread so rapidly making the situation worse. To deal with this dangerous situation and to minimize the further spread of this deadly virus, many countries were locked down and thus even people who are not infected by this virus or not been exposed to this disease have to Self-quarantine to keep themselves safe from getting infected by Corona Virus. The intensity of a lockdown

depends on the situation in which it is declared. In case of COVID-19 pandemic situation, many countries including India declared a complete lockdown. During a complete lockdown period, people are informed to stay where they are and are not allowed to leave their premises. Many organizations started practising Work From Home. People who are dependent on daily wages are deeply affected due to the COVID-19 lockdown situation. Lockdown has adverse effects on the economy, human life, environment and transport sector of the country that in turn leads to unemployment, inflation and recession [1]-[3]. Thus, lockdown disturbs normal life of people. This adversely affects the psychological well-being of people. There are various researchers who have studied the impact of lockdown on the psychological wellbeing of the human being. The researcher discusses the impact of large-scale quarantine during the early 2003 outbreak of severe acute respiratory syndrome (SARS). The research focuses on the factors that influenced people's willingness to follow quarantine orders [4]. Reynolds and Melanie studied the problems, compliance, and psychological impact of the quarantine experience during the SARS pandemic, and the findings imply that quarantine implementation should be evaluated [5],[6]. Researchers have also studied the psychological impacts of quarantining a city in a review study [7], [8]-[13]. Residents in afflicted areas are socially shunned, face workplace discrimination, and have their property vandalized, according to the article. The author of this research studies "coping with the psychological impact of quarantine".The researcher has also explained how quarantine affects mental health, what are the factors that influence coping, and various ways of dealing with the effects of quarantine. [14]. The author developed a Happiness Index survey tool to measure happiness, wellbeing, as well as features of sustainability and resilience. It can also be used to assess happiness with one's life and living circumstances. Survey Development, Domain and Question Reduction, Survey Standardization, and Survey Honing were the four stages of development for the survey instrument [15]. The research on a mental health survey of the UK population before and during the COVID-19 pandemic. The authors find that being young, a woman, and living with children, particularly preschool-aged children, had a significant impact on the extent to which mental anguish rose during the

pandemic[16]. The impact of assessing the prevalence of depression, anxiety, and mental well-being before and during the COVID-19 pandemic is explored in this work by the author [17]-[19]. A review study on “The psychological impact of quarantine”. The authors of this review find that quarantine has a wide-ranging, significant, and long-lasting psychological impact [20]. The author of this research investigated whether or not being quarantined to stop the spread of H1N1 virus had negative psychological impacts [21].

Thus, the present study is about analysing the state of mind of self-quarantined people by measuring their Happiness, identifying the factors that are statistically significant to happiness and to evaluate their sentiments towards quarantining. This study may help policy makers to decide on measures to cut down the psychological consequences of quarantine and to provide guidelines on what things are to be done to take care of the mental health of those who are undergoing quarantining.

## II. METHOD

The current study has followed the methodology as displayed in Fig. 1

### A. Data Collection and Preparation

The data for this study was collected through a survey method. A structured questionnaire designed and shared with people through online mode during second phase (15<sup>th</sup> April 2020 – 3<sup>rd</sup> May 2020) and third phase (4<sup>th</sup> May 2020 – 17<sup>th</sup> May 2020) of COVID-19 lockdown period in India. For this analysis, samples are collected using Stratified Sampling Method. Stratified Sampling is a type of Probability Sampling. In Stratified Sampling method, the population is divided into strata or subgroups and a random sample is taken from each strata [22]. The structured questionnaire designed for this analysis is shared with people who belong to age group 22 and above. Three age groups are created as 22-40, 41-55, 56 & above. A total of 473 responses to the questionnaire are received. After data collection, the data preparation was done. Data preparation includes the data selection and data cleaning. The happiness score will be calculated separately for respondents staying with family and respondents staying away from their family during lockdown period as some parameters that will be used for calculating the happiness score will be different for these two groups of respondents. Thus the data is divided into two datasets. The variables containing text data are used for Sentiment Analysis. The collected data consisted variables such as ‘timestamp’, ‘name’ that were not required for the analysis. Thus, these variables were dropped from the dataset. The null values were not present in the dataset as all the questions were marked as mandatory.

### B. Exploratory Data Analysis (EDA)

The dataset consists of categorical data. To understand how frequently categories of each variable are occurring, frequency distribution technique is used. ‘countplot()’ function of seaborn library is used for plotting the graphs for better understanding of the distribution of categories. From these plots, the categories with very low frequencies will be

identified. The variable with very low frequency categories will be dropped from the further analysis.

### C. Multiple Correspondence Analysis (MCA)

MCA is a method which is usually used to analyse data acquired through a survey questionnaire [23]. The dataset under study consists of many categorical variables. In this study, instead of using correspondence analysis (CA) that is suitable when there are only two categorical variables, MCA is used to understand the relationships between more than two categorical variables. By performing MCA, the similarities between respondents will be identified based on their category selection pattern. From MCA results, the variables that are contributing the most to define dimensions are identified.

### D. Happiness Score Evaluation

To evaluate the happiness score the dataset was divided into two parts as , People who are staying with family (with family dataset) and People who are staying away from their family (without family dataset) . To measure the happiness, the 5-point Likert Scale along with the score 1-5 was used. Table I shows the responses for indicators and their respective score which were used for the analysis. Table II shows the responses for indicators for Yes/No indicators and their score. For measuring happiness of the people who are staying with family and people who are staying away from family, 11 indicators are used. Thus, the maximum score is 55 (11\*5). The happiness score for each respondent is calculated by taking the sum of the score of each indicator.

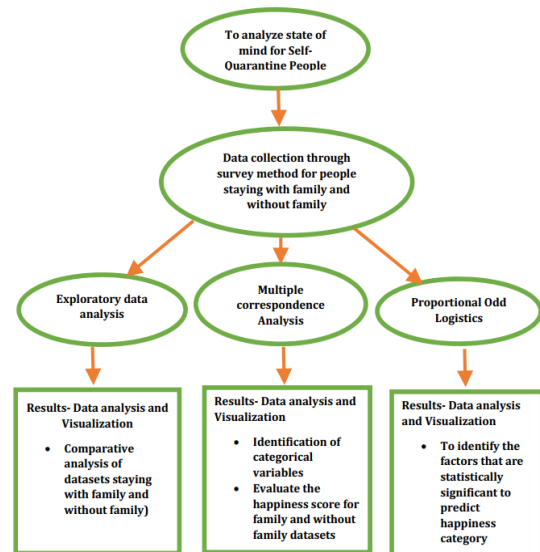


Fig. 1. Research Methodology.

TABLE I. RESPONSE CONVERSION TO SCORE FOR 5 POINT SCALE INDICATORS

Response	Score for ‘sleep’, ‘Governance’ and ‘creativity’	Score for ‘boredom’, ‘discomfort’
0-1	1	5
2-3	2	4
4-6	3	3
7-8	4	2
9-10	5	1

TABLE II. RESPONSE CONVERSION TO SCORE FOR 2 POINT SCALE INDICATORS

Response	Score for 'break needed', 'new skill', 'new routine', 'home exercise',	Score for 'recession', 'feeling lonely', 'fear', 'anxiety', 'stress'
Yes	5	1
No	1	5

GNHI (Gross National Happiness Index) methodology, respondents are classified into one of the four categories as follows (Table III) [24].

TABLE III. HAPPINESS CATEGORIES

Happiness Category	Score Range
Deeply Happy	77% - 100% of the maximum score
Extensively Happy	66% - 76% of the maximum score
Narrowly Happy	50% - 65% of the maximum score
Unhappy	0% - 49% of the maximum score

Thus, for example, if the happiness score of the respondent is 42 out of 55 (77%-100% range) then the respondent is classified as Deeply Happy.

E. Proportional Odds Logistic Regression

The objective for using proportional odds logistic regression model is to identify the factors that are statistically significant to predict happiness category (unhappy/narrowly happy/extensively happy/deeply happy). Based on these factors the happiness category (unhappy/narrowly happy/extensively happy/deeply happy) to which the individual respondent belongs will be predicted. This prediction is done based on the factors that are used to evaluate the happiness score. Here, the happiness categories are ordered and thus proportional odds logistics regression method is used. The proportional odds model can be mathematically represented as:

$$\text{logit} [P (Y \leq j)] = \alpha_j - \sum \beta_i X_i \tag{1}$$

Where, j ranges from 1 to J-1

- Here, J refers to the number of categories of the target variable, in this case the happiness category. Since there are four categories, J = 4.
- The happiness categories are coded as unhappy = 1, narrowly happy = 2, extensively happy = 3 and deeply happy = 4. The category 'deeply happy' is the highest category and 'unhappy' is the lowest category.
- $P(Y \leq 2)$  refers to the probability of being unhappy or narrowly happy versus being extensively happy and above category (in this case, deeply happy).
- Logit refers to 'log odds'. Odds can be defined as the ratio of the probabilities of success of an event and failure of an event. Logit  $[P (Y \leq 1)]$  refers to log odds of the probability [25].
- For better understanding, log odds are converted to probability as follows:

$$P (Y \leq j) = \exp (\alpha_j - \sum \beta_i X_i) / (1 + \exp (\alpha_j - \sum \beta_i X_i)) \tag{2}$$

- Model fitting is done using 'polr()' function from 'MASS' package. The summary result of the model provides intercepts, coefficients of regression with values (slopes), and p-values is calculated from this result [25].
- The coefficients (Variables used in the model) with p-value less than or equal to 0.05 are kept in the model.
- Once the model is finalised, prediction is done on the new values. The 'predict()' function returns estimated probability values for all four categories. The category with highest probability is the category predicted for a respondent [25].

III. RESULTS AND DISCUSSION

A. EDA Results

The total number of respondents staying with their family is 422. The total number of respondents staying away from their family is 51. The frequency distribution of different variables is as follows.

The Fig. 2 and Fig. 3 shows that more than 50% of the respondents of both 'with family' and 'without family' dataset wanted such break from their regular routine that they got due to COVID-19 lockdown. More than 50% of the respondents of both 'with family' and 'without family' dataset planned a new routine for the self-quarantined period (Figure 1 & 2). This shows that instead of being worried, people are trying to adjust with the quarantining situation. It is also shows that, for 'with family' dataset, 55% of the respondents are not learning any new skills whereas for 'without family' dataset, more than 55% of the respondents are learning new skills. This shows that, those who are not staying with their families during self-quarantining period are trying to deal with the loneliness and utilizing the free time they got.

In the questionnaire, respondents were asked to rate the quality of sleep they are getting during self-quarantining period. It is observed that the maximum respondents selected categories for quality of sleep as 7, 8, 9, 10 under both the datasets. Categories 7-9 were considered as, having a good quality sleep while category 10 is considered as having an excellent quality sleep.

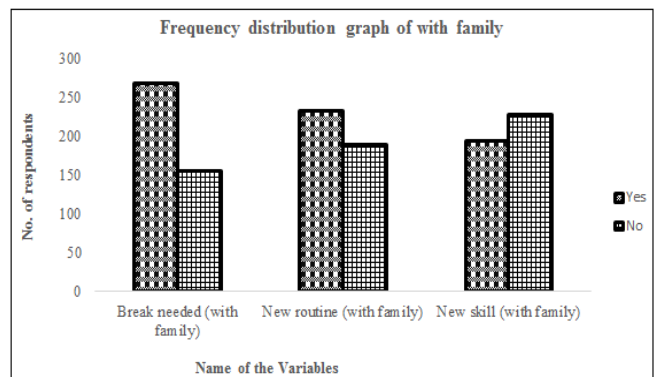


Fig. 2. Response of Indicators (Break Needed, New Routine and New Skill) with Family.



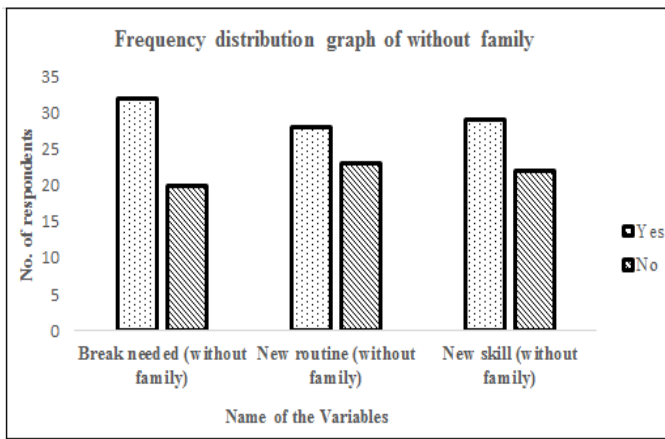


Fig. 3. Response of Indicators (Break Needed, New Routine and New Skill) without Family.

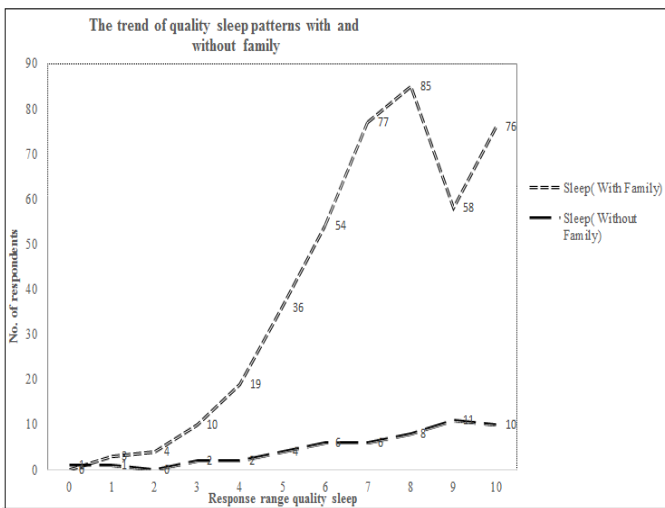


Fig. 4. Quality of Sleep with and without Family.

Around 70% of the total respondents from both the datasets are getting good to excellent quality sleep. Having a good quality sleep helps in maintaining psychological well-being (Fig. 4).

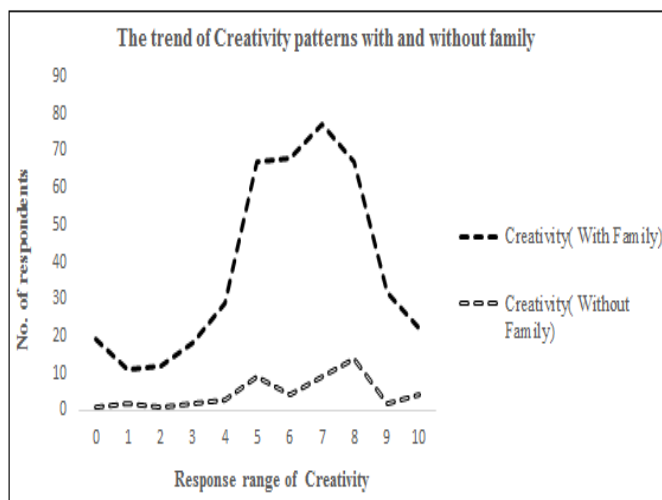


Fig. 5. Creativity with and without Family.

With reference to the creativity attribute, Fig. 5 shows that, most selected categories for ‘with family’ dataset are 5-8 and for ‘without family’ dataset are 5-8 and 10. This indicates that people are trying to be creative to deal with boredom no matter whether they are staying with their family or not.

The Fig. 6 shows the response of boredom from the respondents on a scale of 0-10 being ‘extremely bored’. This question received mixed responses for all the categories (category ‘5’ being the most selected) from the respondents who were staying with their family. Respondents who are staying away from their family experienced high level of boredom.

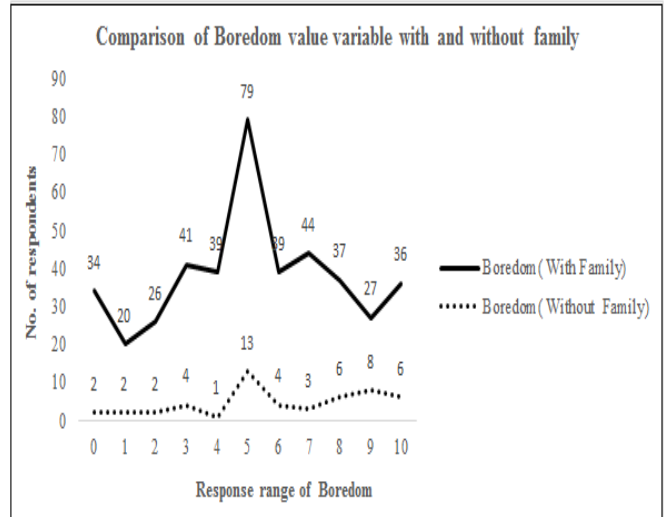


Fig. 6. Comparison of Boredom Value Variable with and without Family.

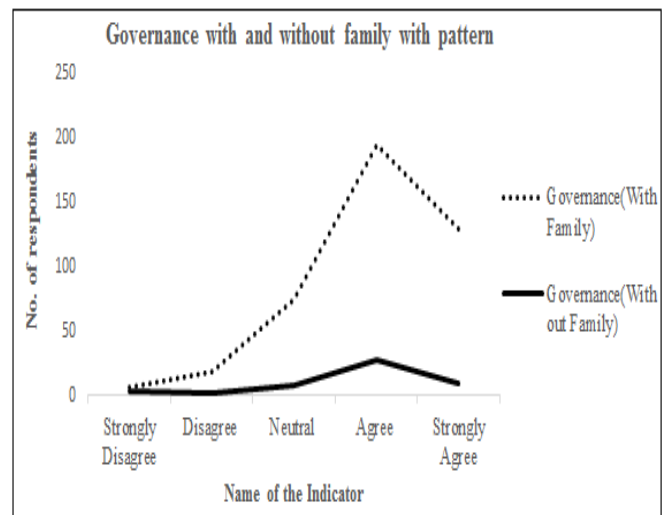


Fig. 7. Comparison Governance with and without Family with Pattern.

Governance is one of the important factor which affects the Happiness of the people. Whether people are satisfied with the strategies adapted by government for handling pandemic is related to the Happiness of population. Fig. 7 shows, the responses of the respondents to the statement ‘The Government is taking the right measures to handle the COVID-19 pandemic situation’. From the graphs it is



observed that from ‘with family’ dataset, 44% of the respondents ‘agree’ and 30% of the respondents ‘strongly agree’ with the above-mentioned statement. From ‘without family’ dataset, 55% of the respondents ‘agree’ and 21% of the respondents ‘strongly agree’ with the above-mentioned statement.

Fig. 8 and Fig. 9 show that, 76% of the respondents from ‘with family’ dataset and 82% of the respondents from the ‘without family’ dataset are worried about the consequences they have to face due to the upcoming recession/inflation. This worry- “about the future”, may affect the psychological well-being of respondents. People may experience anxiety, fear, irritability, stress, depression, fatigue, sadness, panic during quarantining that can impact mental health. From ‘with family’ dataset 62% of the respondents and from ‘without family’ dataset 69% of the respondents experience such feelings (Fig. 8 and Fig. 9). One of the important measures to be taken to stay physically and mentally healthy is to do a regular exercise. Due to lockdown people cannot go to parks or gym. But they can definitely do exercise at home. They can do yoga, meditation, Zumba, home gym etc. Fig. 8 and Fig. 9 shows that, 67% of the respondents of ‘with family’ dataset and 71% of the respondents from ‘without family’ dataset follow home exercise routine.

From ‘with family’ dataset, 95% of the respondents selected category ‘Yes’ for answering whether being with family is helping them to cope with the quarantining situation or not. Also, to answer whether they are satisfied with the quality time spent with their family, 95% of the respondents selected category ‘Yes’ (Fig. 10).

Respondents who are staying with their family during quarantining period may experience discomfort due to extended lockdown (too much togetherness). In response to the statement ‘Extended lockdown (too much togetherness) may cause discomfort with family’, 32% of the respondents selected ‘Disagree’, 29% of the respondents selected ‘Neutral’, 23% of the respondents selected ‘Strongly Disagree’, 13% of the respondents selected ‘Agree’ (Fig. 11).

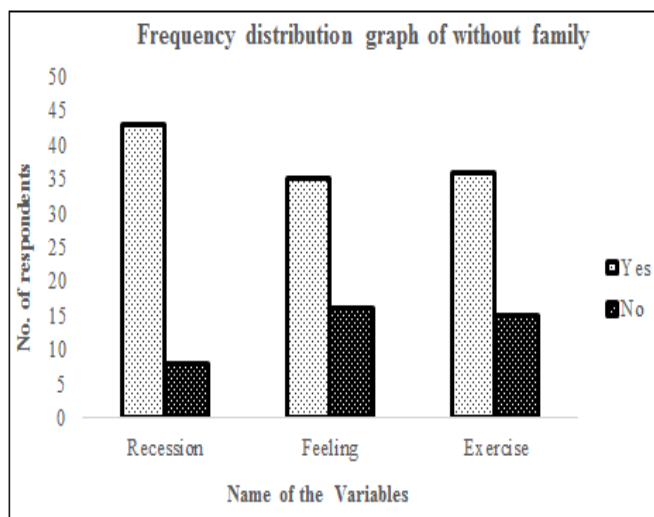


Fig. 9. Response of Indicators (Recession, Feeling, & Exercise) without Family.

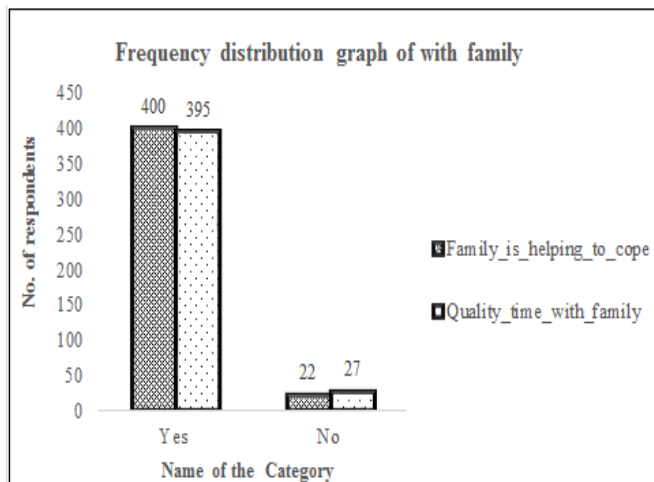


Fig. 10. Response of Indicators (Family\_is\_Helping\_to\_Cope, & Quality\_Time\_with\_Family) with Family.

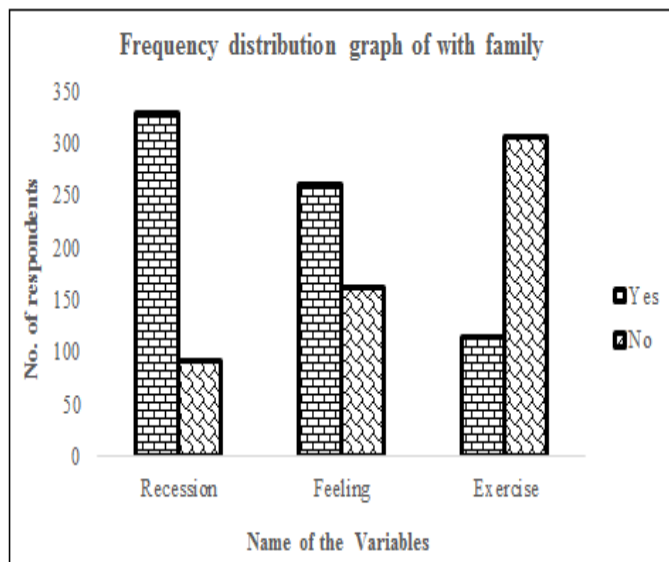


Fig. 8. Response of Indicators (Recession, Feeling, & Exercise) with Family.

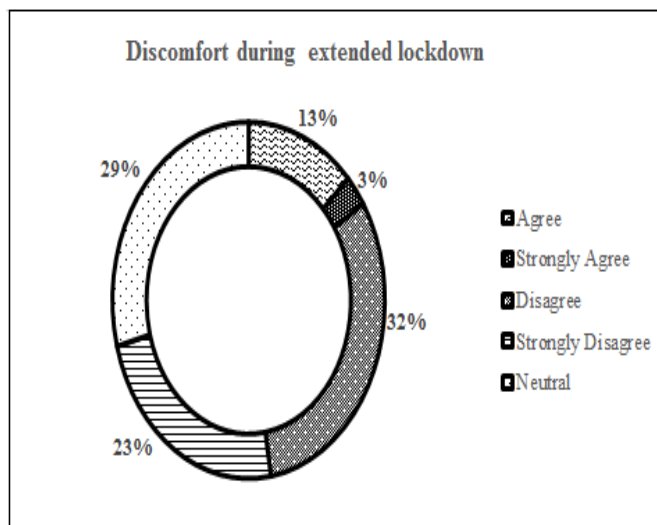


Fig. 11. Response Response of Indicator Discomfort with Family.

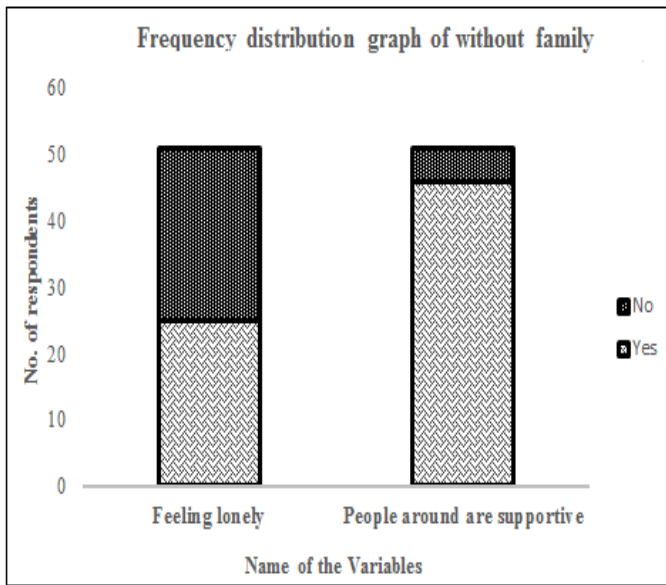


Fig. 12. Response Response of Indicator (Feeling Lonely, & People Around are Supportive) without Family.

Respondents who are not staying with their family during quarantining period may experience loneliness. The question ‘Do you feel lonely as you are away from your family during this self-quarantined situation?’ almost equal number of responses for categories ‘Yes’ and ‘No’. In such difficult situations, getting social support is very important especially for those who are staying away from their family. Respondents staying away from their family were asked whether people around them are supportive or not. In response to this question, 92% of the respondents selected category ‘Yes’. From the above frequency distribution plots, it is found that the variables ‘family is helping to cope’, ‘quality time with family’ from ‘with family’ dataset contain categories having very low frequencies as compared with other category of these variables. Thus, these two variables are omitted from the analysis. Also, the variable ‘people around are supportive’ from ‘without family’ dataset contain category (‘No’) having very low frequency as compared with the other category (‘Yes’) of variable. Thus, this variable is omitted from the analysis (Fig. 12).

**B. MCA Results**

In Fig. 13 and Fig. 14, each point on a graph represents the contribution of that particular variable in constructing dimension one and dimension two. From these graphs we can say that for ‘with family’ dataset, creativity contributes the most in constructing dimension one and dimension two. Whereas for ‘without family’ dataset creativity contributes the most in constructing dimension one and dimension two while new\_skill contributes in constructing dimension one and Governance contributes the most in constructing dimension two.

From the Fig. 15 and Fig. 16 the important categories of the variables were identified. For ‘with family’ dataset, categories sleep\_1, sleep\_2, Governance\_2 contributes the most towards

positive direction of the first dimension whereas categories creativity\_1, discomfort\_with\_family\_1 contributes the most towards positive direction of the second dimension. For ‘without family’ dataset, categories sleep\_1, creativity\_2 contributes the most towards positive direction of the first dimension whereas category Governance\_2 contributes the most towards positive direction of the second dimension. The value of cos2 represents the quality of representation of variables and variable categories. The Fig. 17 and Fig. 18 show that for ‘with family’ dataset, feeling\_1, feeling\_5, governance\_5, exercise\_5, exercise\_1, new\_routine\_1, new\_routine\_5 these categories have higher values of cos2 as compared with other categories for dimension one.

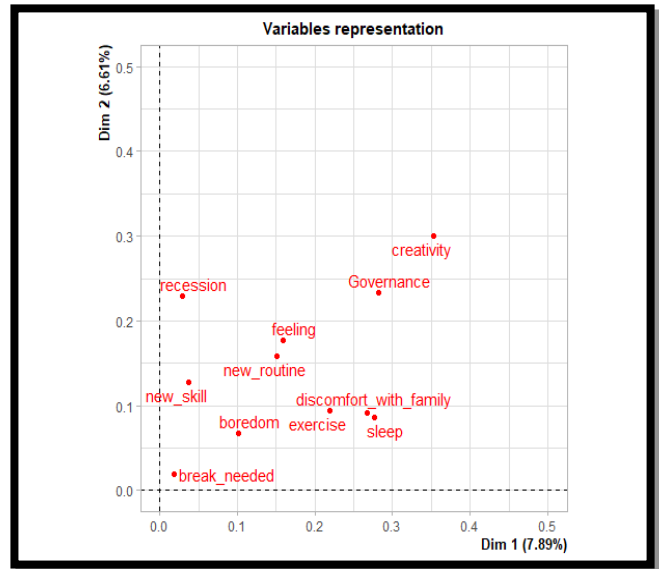


Fig. 13. Variable Representation (with Family).

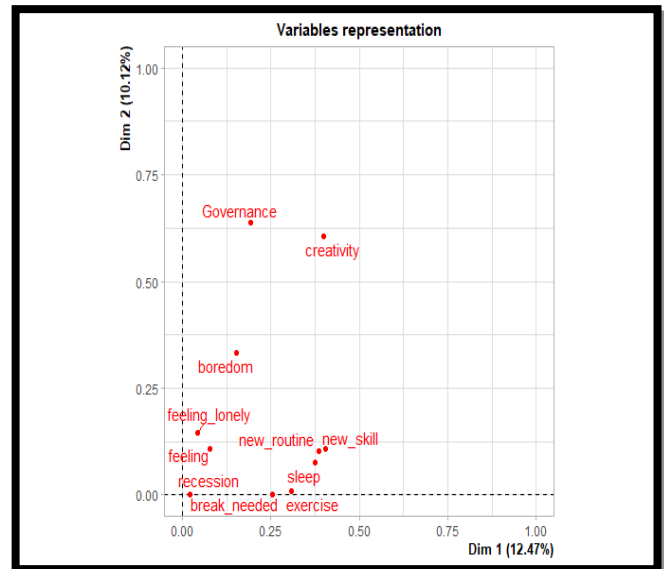


Fig. 14. Variable Representation (without Family).

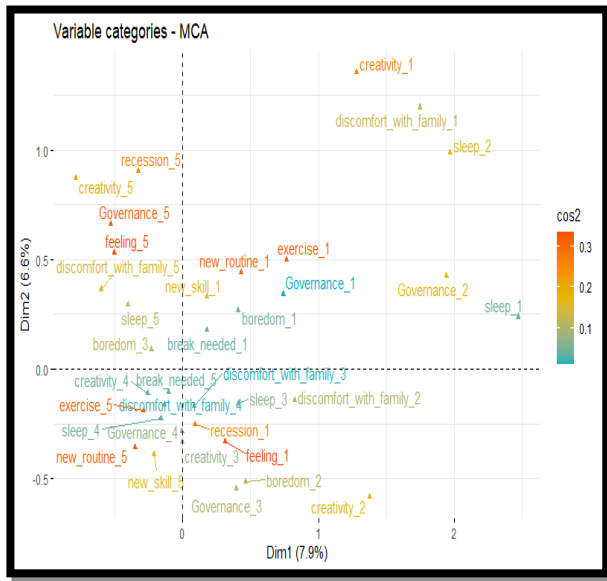


Fig. 15. Variable Categories Plot (with Family).

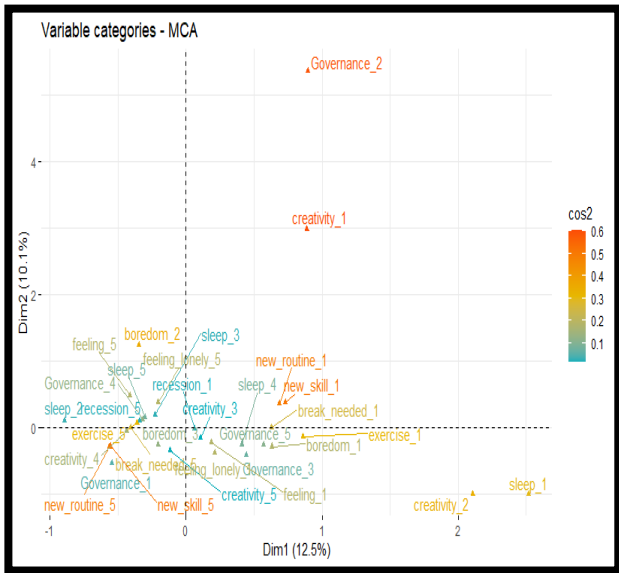


Fig. 16. Variable Categories Plot (without Family).

Whereas, for ‘without family’ dataset, categories creativity\_1, Governance\_2, new\_skill\_5, new\_skill\_1, new\_routine\_1, new\_routine\_5 have higher values of cos2 as compared with other categories for dimension one.

In Fig. 19 and Fig. 20 each point on a graph represents an individual respondent. Respondents are grouped together based on their category selection pattern. From the above graphs it is identified that there are similarities among respondents in both the datasets.

### C. Happiness Score Evaluation Result

Happiness Scores of 422 respondents who are staying with their family is evaluated. Table IV shows that, 15.88 % of

respondents are ‘Deeply Happy’, 34.6 % of the respondents are ‘Extensively Happy’, about 37 % of the respondents are ‘Narrowly Happy’ and 12.56 % of respondents are ‘Unhappy’. For respondents staying without family, we can say that, 17.65 % of respondents are ‘Deeply Happy’, 21.57 % of respondents are ‘Extensively Happy’, 41.17 % of respondents are ‘Narrowly Happy’, 19.6 % of the respondents are ‘Unhappy’. During quarantining situations, factors such as working from home, gender, personality, staying with family (or not) may related to happiness. To verify this, statistical analysis was performed.

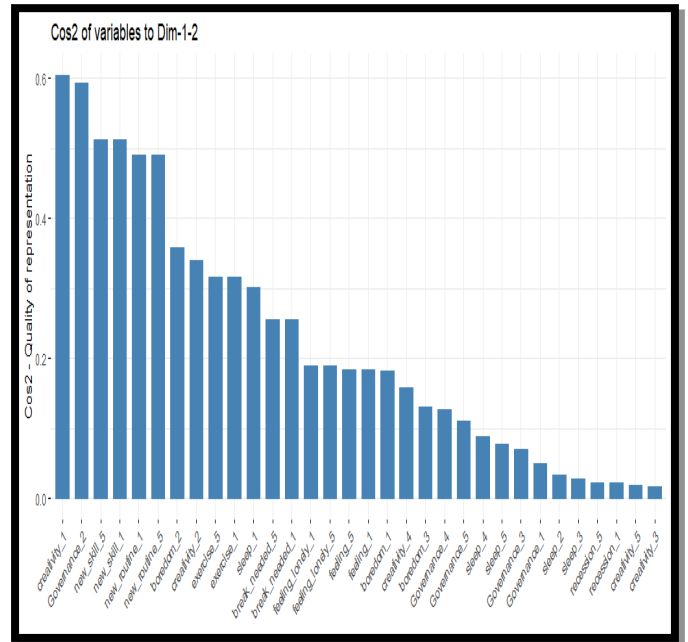


Fig. 17. Cos2 of Variables (with Family).

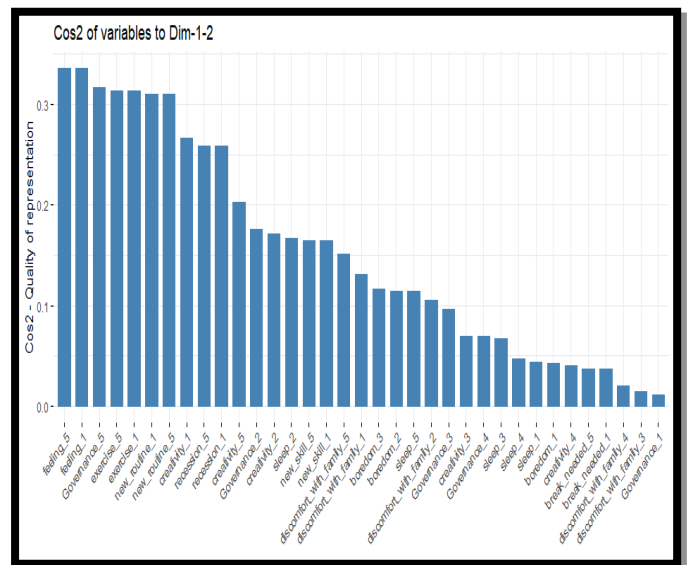


Fig. 18. Cos2 of Variables (without Family).

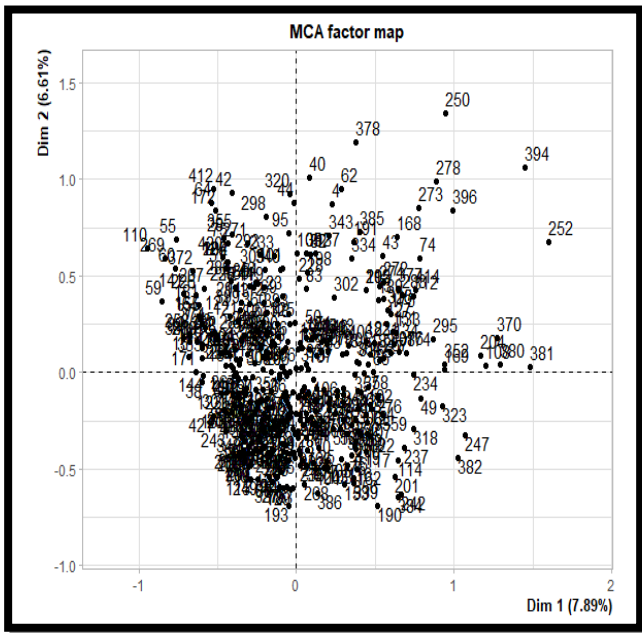


Fig. 19. Individuals Plot (with Family).

are working from home and staying with their family (51%, unhappy and narrowly happy) tend to be happier than respondents who are working from home and staying away from their family (58%, unhappy and narrowly happy).

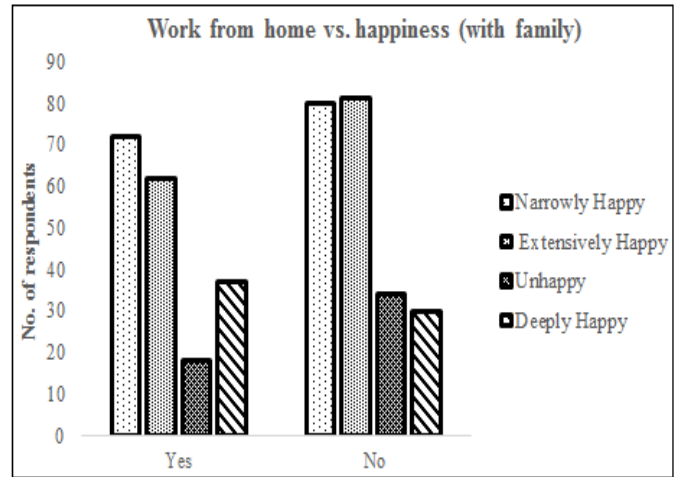


Fig. 21. Work from Home vs. Happiness (with Family).

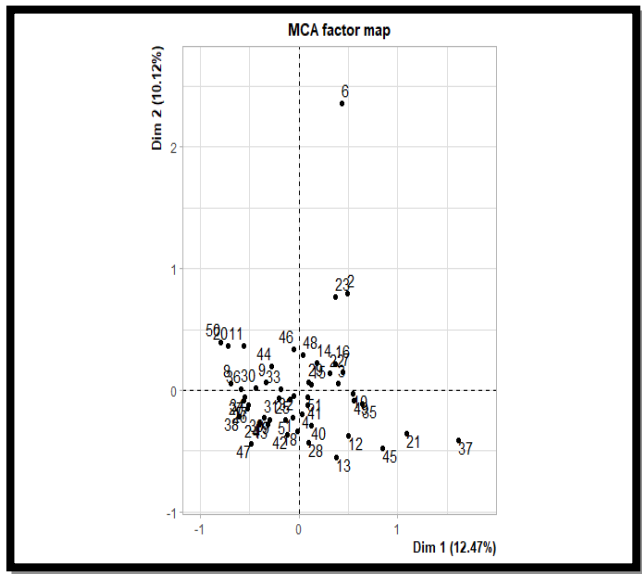


Fig. 20. Individuals Plot (without Family).

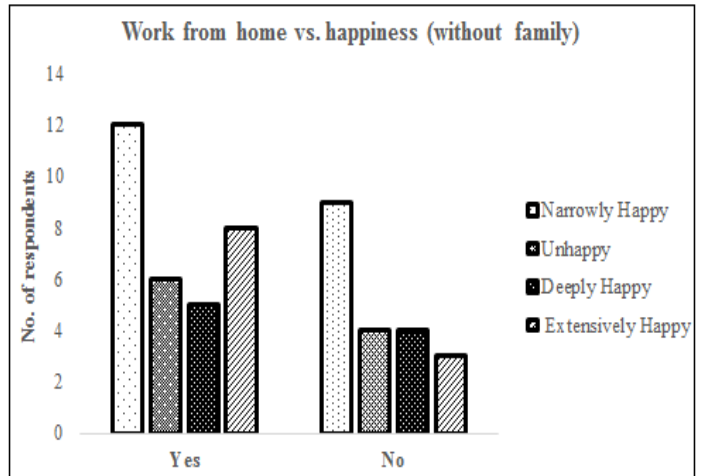


Fig. 22. Work from Home vs. Happiness (without Family).

TABLE IV. HAPPINESS SCORE

Happiness Score	Count of respondents staying with family	Count of respondents staying without family
Deeply Happy	67	9
Extensively Happy	146	11
Narrowly Happy	156	21
Unhappy	53	10
Total	422	51

1) *Work from home vs. happiness*: From the statistical analysis (Fig. 21 and 22) it is found out that respondents who

2) *Gender vs. Happiness*: From Fig. 23 and Fig. 24, it is observed that male respondents staying away from their family tend to be unhappier (57%, unhappy and narrowly happy) than the male respondents staying with their family (47%, unhappy and narrowly happy).

3) *Personality vs. Happiness*: Fig. 25 & Fig. 26 shows that, there is no relation found between personality (Introvert/Extrovert) of the respondent and happiness. But both Extroverts and Introverts who are staying with their family are happier than those who are staying away from their family.

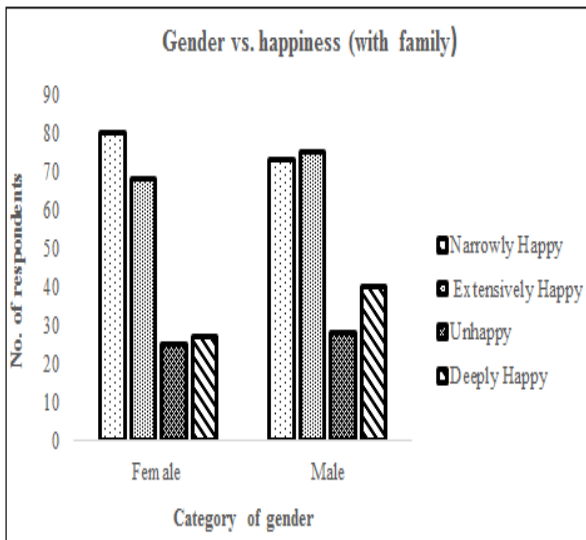


Fig. 23. Gender vs. Happiness (with Family).

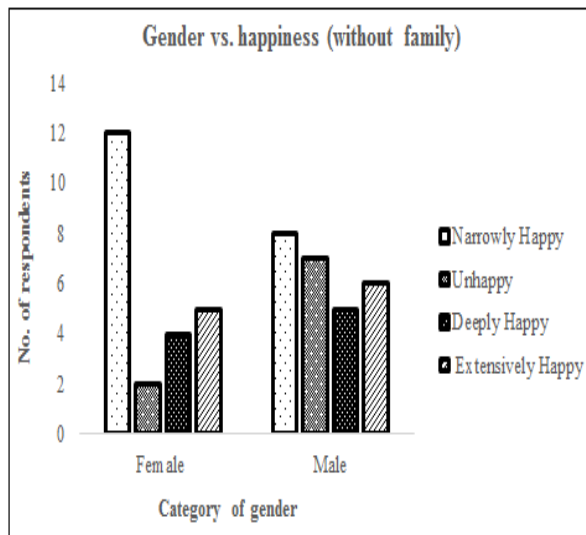


Fig. 24. Gender vs. Happiness (without Family).

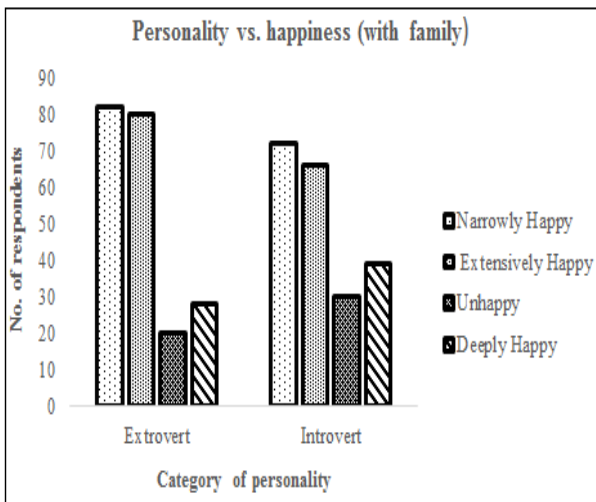


Fig. 25. Personality vs. Happiness (with Family).

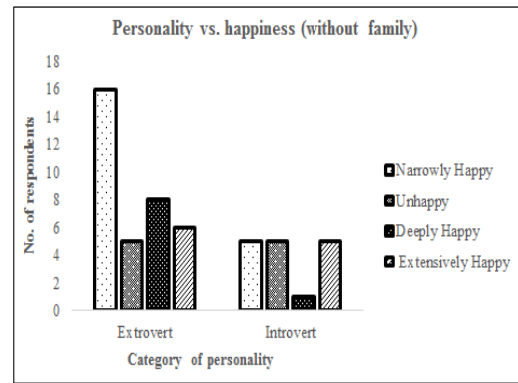


Fig. 26. Personality vs. Happiness (without Family).

D. Proportional Odds Logistic Regression Result

Proportional odd logistic regression model is created using polr() function. Initially all 11 variables that are used to calculate happiness score are included in model creation. In each iteration, the variables with p-value greater than 0.05 are excluded from the model creation (Table V). Finally, a model is created with seven important variables that are break needed, new routine, new skill, recession, discomfort, exercise, feeling for ‘with family’ dataset and with five important variables that are break needed, new routine, new skill, recession, feeling for ‘without family’ dataset (Table VI).

The summary result of the model created for with-family dataset is as follows:

TABLE V. PROPORTIONAL ODD LOGISTIC REGRESSION SUMMARY RESULT (WITH FAMILY)

Coefficients	Value	Std. Error	t-value	p-value
break_needed5	4.563	0.4799	9.507	0.000
new_routine5	4.311	0.4611	9.350	0.000
new_skill5	4.563	0.4764	9.577	0.000
recession5	4.235	0.4812	8.801	0.000
discomfort2	3.151	0.9342	3.373	0.001
discomfort3	5.199	0.9549	5.444	0.000
discomfort4	5.743	0.9567	6.003	0.000
discomfort5	6.945	1.0082	6.888	0.000
exercise5	4.972	0.5193	9.573	0.000
feeling5	5.342	0.5262	10.152	0.000

TABLE VI. TABLE VI. INTERCEPTS (WITH FAMILY)

Proportion	Value	Std. Error	t-value	p-value
1 2	11.2652	1.2860	8.7600	0.000
2 3	19.3984	1.8702	10.3722	0.000
3 4	25.9091	2.3327	11.1069	0.000

TABLE VII. TABLE VII. RESIDUAL DEVIANCE AND AIC (WITH FAMILY)

Residual Deviance	382.3822
AIC	408.3822



Interpretation:

- The coefficient values correspond to  $\beta$  and intercept values correspond to  $\alpha$ .
- Residual Deviance and AIC values are used to compare different models.
- The categorical variables can be interpreted as follows:
- Variable 'new\_skill' can be interpreted as the respondent who selected category of new\_skill variable as '5' tends to be happier than the respondent who selected category '1'. Other variables are interpreted in the same manner.
- Intercept interpretation:

Table VII shows that the intercept 1|2 (i.e Unhappy|Narrowly Happy) corresponds to  $\text{logit}[P(Y \leq 1)]$ . It can be interpreted as log odds of being 'Unhappy' versus being 'Narrowly Happy', 'Extensively Happy' or 'Deeply Happy'.

The intercept 2|3 (i.e Narrowly Happy|Extensively Happy) corresponds to  $\text{logit}[P(Y \leq 2)]$ . It can be interpreted as log odds of being 'Unhappy' or 'Narrowly Happy' versus being 'Extensively Happy' or 'Deeply Happy'.

The intercept 3|4 (i.e Extensively Happy| Deeply Happy ) corresponds to  $\text{logit}[P(Y \leq 3)]$ . It can be interpreted as log odds of being 'Unhappy' or 'Narrowly Happy' or 'Extensively Happy' versus being 'Deeply Happy'.

The summary result of the model created for without-family dataset is shown in the Table VIII.

The interpretation of this result is same as that of with-family dataset result Table IX and Table X.

The prediction function returns the estimated probabilities for each class (Unhappy/Narrowly Happy/Extensively Happy/Deeply Happy)

For with-family dataset the prediction result is as follows:

For data values break\_needed = 5, new\_routine = 1, new\_skill = 1, recession = 1, discomfort = 2, exercise = 5, feeling =5 the total score is 20 out of 35 (57.14 %) the prediction output is shown in Table XI.

The estimated probability for class 2 (Narrowly Happy) is highest. Thus, the model correctly predicts the class for given data values.

Similarly, for without-family dataset the prediction result is shown in the Table XII.

For data values break\_needed = 5, new\_routine =1, new\_skill =5, recession=1, feeling=1 the total score is 13 out of 25 (52 %) the prediction output is:

Validation:

- In the questionnaire, the respondents were asked to describe self-quarantining in one word. To validate the happiness score calculated for each respondent, this

one word mentioned by the respondent is used to compare it with the happiness category.

- To perform validation, happiness categories 'Unhappy' and 'Narrowly Happy' are coded as 0 and happiness categories 'Extensively Happy' and 'Deeply Happy' are coded as 1. The positive words used to describe self-quarantining are coded as 1 and negative words that are used to describe self-quarantining are coded as 0. Then the number of 1's and 0's for both the happiness categories and one word are counted.
- For with-family dataset, number of 0's for one word is 61 and number of 1's is 131. For happiness categories, number of 0's is 75 and that of 1's is 117. Thus, 89.31 % respondents are correctly categorized as 'Happy'. And 81.33 % respondents are correctly categorized as 'Unhappy'.
- For without-family dataset, number of 0's for one word is 8 and number of 1's is 18. For happiness categories, number of 0's is 13 and that of 1's is 13. Thus, 72.22 % respondents are correctly categorized as 'Happy'. And 61.53 % respondents are correctly categorized as 'Unhappy'.

TABLE VIII. PROPORTIONAL ODD LOGISTIC REGRESSION SUMMARY RESULT (WITHOUT FAMILY)

Coefficients	Value	Std. Error	t-value	p-value
break_needed5	2.605	0.7730	3.370	0.001
new_routine5	2.839	0.9018	3.148	0.002
new_skill5	2.422	0.8092	2.993	0.003
recession5	2.178	0.9726	2.239	0.025
feeling5	2.980	0.8189	3.639	0.000

TABLE IX. INTERCEPTS (WITHOUT FAMILY)

Propotion	Value	Std. Error	t-value	p-value
1 2	2.2204	0.7305	3.0396	0.002
2 3	6.7077	1.3324	5.0343	0.000
3 4	9.7029	1.7837	5.4397	0.000

TABLE X. RESIDUAL DEVIANCE AND AIC (WITHOUT FAMILY)

Residual Deviance	70.69671
AIC	86.69671

TABLE XI. PREDICTION OUTPUT (WITH FAMILY)

1	2	3	4
0.001	0.796	0.202	0.000

TABLE XII. PREDICTION OUTPUT (WITHOUT FAMILY)

1	2	3	4
0.057	0.786	0.148	0.009

### E. Discussion

- This study found that respondents who needed break from regular routine, who planned a new daily routine



to follow during quarantine, who utilized the available free time by learning some new skills, who exercise regularly during quarantine to be physically fit, tend to be happier.

- Respondents experiencing anxiety, stress, fear, irritability, frustration, panic tend to be unhappier.
- This study suggested that the family factor is related to the happiness of the respondents.
- Respondents who are staying with their family may feel discomfort due to extended lockdown and this may lead to unhappiness.
- It is also found that the happiness of the respondents who are working from home during the self-quarantining period is related to whether they are staying with their family or not. Also, irrespective of the personality of the respondent, those who are staying with their family during self-quarantining period tend to be happier than those who are staying away from their family.
- Male respondents who are staying away from their family tend to be unhappier.
- The worry about the future consequences such as inflation or recession that the respondents may have to face post lockdown is related to the happiness.
- From the sentiment analysis it is found out that most respondents have a positive attitude towards self-quarantining that leads to respondents being psychologically healthy.

#### F. Limitations

This study has some limitations.

- The sample size is small. Thus, the results are not generalizable.
- The samples are taken during second and third phase of the COVID-19 lockdown period. If samples were taken during fourth phase, then the results might have been different.

#### IV. CONCLUSION

The overall happiness of self-quarantined people is measured. This study identified the factors affecting happiness of those who undergo quarantining. For 'with family' dataset, these factors include 'break needed', 'new routine', 'new skill', 'recession', 'discomfort with family', 'exercise' and 'feeling'. For 'without family' dataset, these factors include 'break needed', 'new routine', 'new skill', 'recession' and 'feeling'. The sentiments of self-quarantined people towards quarantining are evaluated. This study may help to identify the measures that can be taken to mitigate the consequences of Quarantine. According to the results, people who undergo quarantining can be advised to plan a routine, utilize the free time by learning some new skills, to do regular exercise at home. This will keep them happy and mentally healthy which

eventually will help them to cope with the quarantining situation.

#### REFERENCES

- [1] S. Nundy, A. Ghosh, A. Mesloub, G. A. Alabaqawy, and M. M. Alnaim, "Impact of COVID-19 pandemic on socio-economic, energy-environment and transport sector globally and sustainable development goal (SDG)," *J. Clean. Prod.*, vol. 312, p. 127705, 2021, doi:10.1016/j.jclepro.2021.127705.
- [2] S. Abdullah et al., "Air quality status during 2020 Malaysia Movement Control Order (MCO) due to 2019 novel coronavirus (2019-nCoV) pandemic," *Sci. Total Environ.*, vol. 729, p. 139022, 2020, doi:10.1016/j.scitotenv.2020.139022.
- [3] M. Abdullah, C. Dias, D. Muley, and M. Shahin, "Exploring the impacts of COVID-19 on travel behavior and mode preferences," *Transp. Res. Interdiscip. Perspect.*, vol. 8, p. 100255, 2020, doi:10.1016/j.trip.2020.100255.
- [4] C. DiGiovanni, J. Conley, D. Chiu, and J. Zaborski, "Factors influencing compliance with quarantine in Toronto during the 2003 SARS outbreak," *Biosecur. Bioterror.*, vol. 2, no. 4, pp. 265–272, 2004, doi: 10.1089/bsp.2004.2.265.
- [5] D. L. Reynolds, J. R. Garay, S. L. Deamond, M. K. Moran, W. Gold, and R. Styra, "Understanding, compliance and psychological impact of the SARS quarantine experience," *Epidemiol. Infect.*, vol. 136, no. 7, pp. 997–1007, Jul. 2008, doi: 10.1017/S0950268807009156.
- [6] M. R. Taylor, K. E. Agho, G. J. Stevens, and B. Raphael, "Factors influencing psychological distress during a disease epidemic: Data from Australia's first outbreak of equine influenza," *BMC Public Health*, vol. 8, pp. 1–13, 2008, doi: 10.1186/1471-2458-8-347.
- [7] G. J. Rubin and S. Wessely, "The psychological effects of quarantining a city," *BMJ*, vol. 368, p. m313, Jan. 2020, doi: 10.1136/bmj.m313.
- [8] F. J. de Oliveira Araújo, L. S. A. de Lima, P. I. M. Cidade, C. B. Nobre, and M. L. R. Neto, "Impact Of Sars-Cov-2 And Its Reverberation In Global Higher Education And Mental Health," *Psychiatry Res.*, vol. 288, p. 112977, 2020, doi: 10.1016/j.psychres.2020.112977.
- [9] G. J. G. Asmundson and S. Taylor, "How health anxiety influences responses to viral outbreaks like COVID-19: What all decision-makers, health authorities, and health care professionals need to know," *J. Anxiety Disord.*, vol. 71, p. 102211, 2020, doi: 10.1016/j.janxdis.2020.102211.
- [10] S. Galea, R. M. Merchant, and N. Lurie, "The Mental Health Consequences of COVID-19 and Physical Distancing: The Need for Prevention and Early Intervention," *JAMA Intern. Med.*, vol. 180, no. 6, pp. 817–818, 2020, doi: 10.1001/jamainternmed.2020.1562.
- [11] J. Qiu, B. Shen, M. Zhao, Z. Wang, B. Xie, and Y. Xu, "A nationwide survey of psychological distress among Chinese people in the COVID-19 epidemic: Implications and policy recommendations," *Gen. Psychiatry*, vol. 33, no. 2, pp. 1–4, 2020, doi: 10.1136/gpsych-2020-100213.
- [12] R. P. Rajkumar, "COVID-19 and mental health: A review of the existing literature," *Asian J. Psychiatr.*, vol. 52, p. 102066, 2020, doi: 10.1016/j.ajp.2020.102066.
- [13] X. Zhou et al., "The Role of Telehealth in Reducing the Mental Health Burden from COVID-19," *Telemed. e-Health*, vol. 26, no. 4, pp. 377–379, 2020, doi: 10.1089/tmj.2020.0068.
- [14] R. Laslo-Roth, S. George-Levi, and M. Margalit, "Hope during the COVID-19 outbreak: coping with the psychological impact of quarantine," *Couns. Psychol. Q.*, vol. 34, no. 3–4, pp. 771–785, 2021, doi: 10.1080/09515070.2021.1881762.
- [15] L. Musikanski et al., "Happiness Index Methodology," *J. Soc. Chang.*, vol. 9, 2017, doi: 10.5590/JOSC.2017.09.1.02.
- [16] M. Pierce et al., "Mental health before and during the COVID-19 pandemic: a longitudinal probability sample survey of the UK population," *The Lancet. Psychiatry*, vol. 7, no. 10, pp. 883–892, Oct. 2020, doi: 10.1016/S2215-0366(20)30308-4.
- [17] A. S. F. Kwong et al., "Mental health before and during the COVID-19 pandemic in two longitudinal UK population cohorts," *Br. J. Psychiatry*, vol. 218, no. 6, pp. 334–343, 2021, doi: 10.1192/bjp.2020.242.

- [18] M. M. Kshirsagar, A. S. Dodamani, G. A. Dodamani, V. R. Khobragade, and R. N. Deokar, "Impact of Covid-19 on Mental Health: An Overview," *Rev. Recent Clin. Trials*, vol. 16, no. 3, pp. 227–231, 2021, doi: 10.2174/1574887115666210105122324.
- [19] W. Cao et al., "The psychological impact of the COVID-19 epidemic on college students in China," *Psychiatry Res.*, vol. 287, p. 112934, 2020, doi: 10.1016/j.psychres.2020.112934.
- [20] S. K. Brooks et al., "The psychological impact of quarantine and how to reduce it: rapid review of the evidence," *Lancet*, vol. 395, no. 10227, pp. 912–920, 2020, doi: 10.1016/S0140-6736(20)30460-8.
- [21] Y. Wang, B. Xu, G. Zhao, R. Cao, X. He, and S. Fu, "Is quarantine related to immediate negative psychological consequences during the 2009 H1N1 epidemic?," *Gen. Hosp. Psychiatry*, vol. 33, no. 1, pp. 75–77, 2011, doi: 10.1016/j.genhospsych.2010.11.001.
- [22] H. Taherdoost, "Sampling Methods in Research Methodology; How to Choose a Sampling Technique for Research", *International Journal of Academic Research in Management (IJARM)*, vol. 5, no. 2, pp. 18-27, 2016, doi: 10.2139/ssrn.3205035.
- [23] F. Husson and J. Josse, "Multiple Correspondence Analysis", in *Visualization and Verbalization of Data*, Chapman and Hall/CRC, 1st ed. Boca Raton, Florida, US, 2014, pp.163-181
- [24] S. Pratt, "A Gross Happiness Index for the Solomon Islands and Tonga: an Exploratory Study," *Glob. Soc. Welf.*, vol. 3, no. 1, pp. 11–21, 2016, doi: 10.1007/s40609-015-0041-1..
- [25] C. L. Lackner and C. H. Wang, "Demographic, psychological, and experiential correlates of SARS-CoV-2 vaccination intentions in a sample of Canadian families," *Vaccine X*, vol. 8, p. 100091, 2021, doi: 10.1016/j.jvaxc.2021.100091.

# SIBI (Sign System Indonesian Language) Text-to-3D Animation Translation Mobile Application

Erdefi Rakun, Sultan Muzahidin, IGM Surya A. Darmana, Wikan Setiaji

Faculty of Computer Science  
Universitas Indonesia  
Depok, Indonesia

**Abstract**—This research proposed a mobile application prototype to translate Indonesian text into SIBI (Sign System for the Indonesian Language) 3D gestures animation to bridge the communication gap between the deaf and the other. To communicate in sign language, the signer will use his/her hands and fingers to demonstrate the word gesture, and at the same time, his/her mouth will pronounce the word being expressed. Therefore, the proposed mobile application needs two animation generator components: the hand gesture and the lip movement generator. Hand gestures are made using a motion capture sensor. Mouth movements are created for all syllables available in the SIBI dictionary using the Dirichlet Free-Form Deformation (DFFD) method. The subsequent challenging work is synchronizing these two components and adding transitional gestures. A transitional gesture done by the cross-fading method is needed to make a word gesture that can smoothly connect with the next word gesture. The Mean Opinion Score (MOS) test was run to measure the mouth movements in 3D animation. The MOS score is 4.422. There are four surveys conducted to measure user satisfaction. The surveys showed that the animation generated did not significantly differ from the original video. The Sistem Usability Score (SUS) is 76.25. The score means that prototype is in the GOOD category. The average time needed to generate an animation from Indonesian input text is less than 100ms.

**Keywords**—SIBI sign language; sequence generation; visual speech; animation

## I. INTRODUCTION

Sign language is a non-verbal language used to help people with hearing impairment communicate. Sign language represents a word with hand gestures and mouth movements. Communication with sign language uses a combination of hand, finger, and mouth movements representing words [1]. Indonesia has two sign languages acknowledged by the government: SIBI (Sign System for the Indonesian Language) and BISINDO (Indonesian Sign Language).

SIBI is a sign language that the Ministry of Education and Culture officially acknowledged in Indonesia in 1994. SIBI follows the Indonesian language grammar and has been used formally in the School for special needs students. The characteristic of SIBI is that SIBI applies Indonesian grammar

in organizing word gestures in a sentence [2]. Indonesian words are written using the Latin-Roman alphabet and categorized into four elements: subject, verb, noun, and adverb. Indonesian also has inflectional words that attach prefixes, suffixes, and affixes to the root word. With these affixes, the root word has additional meaning.

BISINDO is a sign language that developed naturally through the deaf community in Indonesia. BISINDO does not follow Indonesian grammar and is commonly used in conversation. BISINDO prioritizes the meaning of the gestures carried out rather than the language structure of the gestures.

Unfortunately, not many people master sign language to communicate with the Deaf. This research proposes to bridge the communication gap between the deaf and others by building a mobile application to translate Indonesian text to 3D SIBI gesture animation.

SIBI differs from other sign languages such as American Sign Language (ASL) and British Sign Language (BSL) in terms of their gestures and method of arranging gestures in a sentence. Gestures in SIBI are arranged according to the rules in Indonesian grammar. Another difference lies in how the inflectional gesture is formed. Inflectional gestures are formed by combining the root word and affixes in the inflectional words [3].

Constructing a SIBI sentence gesture that differs from other sign languages needs different ways to generate 3D animation of a SIBI sentence gesture. This research faces several challenges. First, the Indonesian input sentence must be deconstructed into its components according to the SIBI rules to generate animated gestures. Fig. 1 shows an example of how an Indonesian sentence is deconstructed into word components: An inflectional word will be split into components affixes, and the root word ("mengatakan" = claim, will be separated into to prefix "me" + root word "kata" + suffix "kan"); Name will be changed to its alphabets ("William" will be split into w+i+l+l+i+a+m); numbers will be split into their essential number components ("6023" becomes "6"+"thousand"+"20"+3) [4].

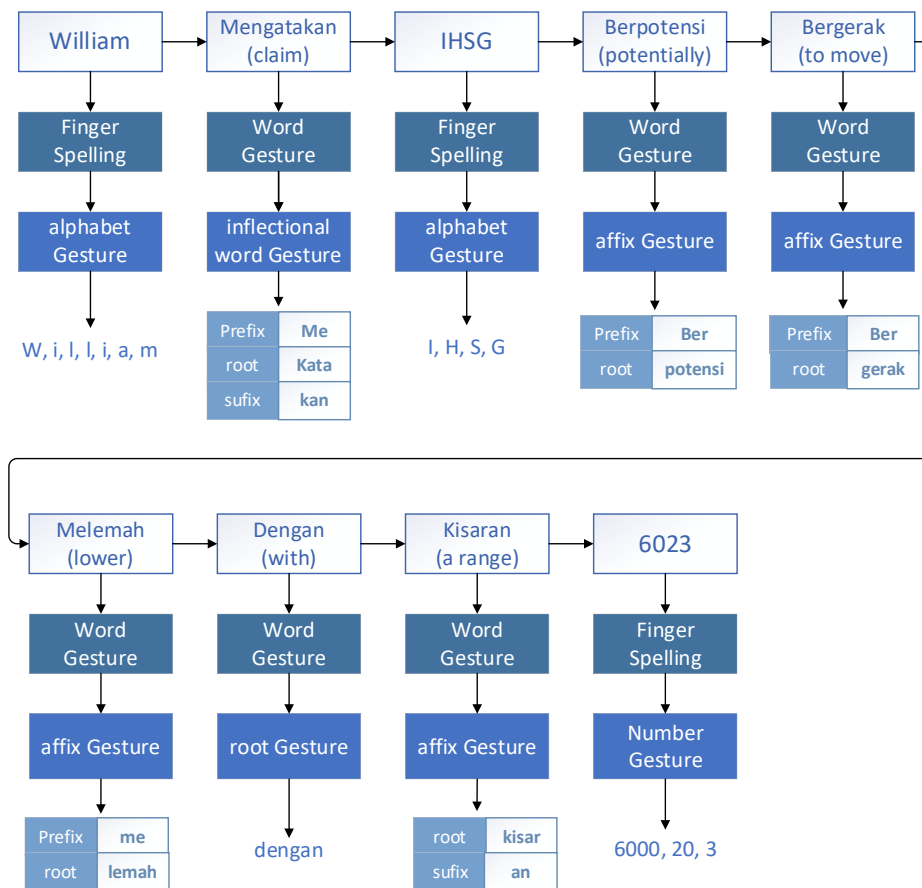


Fig. 1. Deconstruction of an Indonesian Input Sentence.

Second, to communicate in sign language, the signer will use his/her hands and fingers to demonstrate the word gesture, and at the same time, his/her mouth will pronounce the word being expressed. Therefore, building a 3D SIBI gesture animation has to be divided into two steps: building hand gestures and mouth movements. After the hand and mouth movement components are completed, two more challenges need to be solved by this research: how to make hand movements synchronize with mouth movements and how to connect the word components in a sentence into a single, smooth movement.

In conclusion, this paper discusses how to solve the four challenges in building an application to translate Indonesian input text into a 3D animation of the SIBI gesture.

The remainder of this paper is organized as follows: Section 2 states other research on generating text-to-gesture translation systems and mouth movement; Section 3 explains the proposed method for Indonesian text-to-3D SIBI gesture animation, dataset, and evaluation metric; Section 4 evaluates and analyzes the evaluation results; Section 5 closes this paper with the conclusion and future works.

## II. RELATED WORKS

This section discusses other research on generating text-to-gesture translation systems and mouth movement. Table I shows some research on gesture generation, while Table II shows the research on mouth movements.

Table I shows that a common way to generate gestures from text is to create a sign language script and then generate gestures based on the script. Sign language scripts commonly used are Sutton SignWriting and HamNoSys notation. Sutton SignWriting notation is sign language that transcribes the signed gestures spatially, in two-dimensional canvas, as they are visually perceived. Furthermore, HamNoSys notation is an alphabetic system that describes a sign, primarily phonetic. HamNoSys notation is designed as a markup language foundation used to transcribe all sign languages worldwide. It does not depend on the conventions of each country, such as gestures for spelling the finger alphabet [5][6][7]. The lack of documentation of the HamNoSys and Sutton SignWriting corpus available for sign language in Indonesia causes the HamNoSys, and Sutton SignWriting cannot be implemented in SIBI's text-to-gesture translation system. SIBI is a sign system that follows Indonesian grammar [8]. The similarity between the SIBI structure and the Indonesian language structure is an advantage that can be used in SIBI's text-to-gesture translation system [4]. This research proposed an Indonesian Language stemming method to find word components in SIBI sentences. Table I also shows that other research focuses on building web-based text-to-gesture translation systems. Meanwhile, this research focuses on developing a text-to-gesture translation system as an Android Mobile Application.

TABLE I. RESEARCH ON GENERATING GESTURES IN SEVERAL COUNTRIES

Author	Sign Language	Platform	Method
(Karpouzis et al., 2007) [5]	Greek Sign language	Web-Based	1. Scripting Technology for Embodied Personal language (STEP) 2. HamNoSys 3. 3D Animation
(Bouزيد & Jemni, 2014) [28]	Tunisian Sign Language	Web-Based	1. SWML (SignWriting Markup Language) 2. Sutton SignWriting notation 3. 3D Animation
(Boulares & Jemni, 2012) [6]	American Sign Language	Web-Based and Android	1. XML based 2. HamNoSys 3. Video Animation
(Efthimiou et al., 2009)[7]	Dicta-Sign: Greek, British, German, and French Sign Language	Web-Based	1. Signing Gesture Markup Language (SiGML) 2. HamNoSys 3. 3D Animation

Table II shows the research on mouth movements. Mouth movement research usually uses the form of viseme, which is a visual form of pronunciation. Three approaches to generating Viseme derived automatically from pronunciation are key-frame interpolation, model-based, and concatenative [9]. The Key-frame interpolation connects through interpolation the pre-define lip shape of all viseme that appear in a word or sentence [10], [11]. The model-based approach creates viseme from each pronunciation done by the human model [12]. The concatenative approach is a combination of key-frame and model-based approaches. Research by [13] and [14] tracks viseme on the human face to create a database of viseme animations. To generate mouth movement animation, all the viseme that appear in a word or sentence will be taken from the database and then connected through interpolation. The concatenative approach creates a realistic speech animation because it uses actual human face data as a model. Therefore, this research uses the concatenative approach for SIBI mouth movement animation.

Research related to text-to-gesture translation systems generally focuses on generating hand movements only. So far, no system has been found to generate hand movements and mouth movements from text input. This research proposes a combined SIBI hand and mouth movements from Indonesian text.

TABLE II. THE GENERATION OF MOUTH MOVEMENTS RESEARCH

Author	Language	Input	Method
(Setyati et al., 2017) [10]	Indonesian	Text	Hidden Markov Model, 2D Animation, key-frame interpolation
(Haryanto & Sumpeno, 2018) [11]	Indonesian	Text	Morphing Viseme, Syllable Concatenation, FACS, Key-frame Interpolation
(Yu & Wang, 2015) [12]	Mandarin Chinese	Video, Voice, dan Text	AAM, RBF Interpolation, Model-Based
(Ni & Liu, 2019) [13]	Chinese	Voice	DFFD, Concatenative
(Taylor et al., 2017) [14]	English	Voice	AAM, Concatenative

Note: FACS = Facial Action Coding System, AAM = Active Appearance Model, RBF = Radial Basis Function, DFFD = Dirichlet Free-Form Deformation

### III. PROPOSED METHOD

This section discusses the proposed method to generate 3D animation from Indonesian sentence text. The discussion is divided into six sub-sections: overall system design, how to do SIBI sentence deconstruction, how to make 3D animation for hand and mouth movements, how to synchronize hand and mouth movements, and how to connect each word component in sentences by inserting transitional movements, and evaluations carried out to measure system performance.

#### A. Application Overview

Fig. 2 is the architecture used to build the Indonesian text to SIBI's 3D animation. The application consists of two main modules: text parser and animation engine. Text parser consists of two types: deconstructing Indonesian sentences into word components using hand gesture text parser and deconstructing Indonesian sentences into syllables using mouth movement text parser.

The other process is to generate movement based on the text parser. This process occurs in the animation engine module, resulting a 3D animation. This module is divided into two, namely, hand gestures and mouth movement animation engines. The hand gesture animation engine develops 3D animation of hand and finger movements based on data obtained from sensors placed on the body of the SIBI expert. Meanwhile, the mouth movement animation engine develops mouth movements using the facial data of the SIBI expert model. The two-generation processes will run in parallel. The synchronization process equalizes the speed between the hand gesture and mouth animation movements.

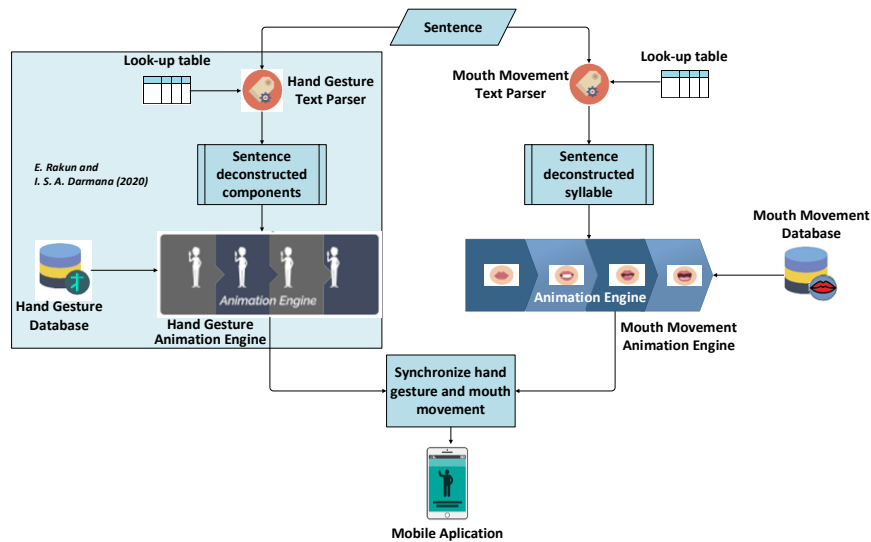


Fig. 2. SIBI 3D Animation Application Architecture.

The previous study (Rakun & Darmana 2020) [15] discusses deconstructing Indonesian sentences into word components and the hand gesture animation engine, so these two topics are discussed briefly here. This paper will discuss the mouth movement text parser, the mouth movement animation engine, and how to synchronize hand and mouth animations.

**B. SIBI Sentence Deconstruction**

This section discusses breaking down Indonesian sentences into the components needed to generate 3D animation for hand movements (1) and mouth movements (2). The text parsing result of both hand and mouth in (3).

1) *Hand gesture text parser*: SIBI uses the standard Indonesian grammar and has two types of gestures: word and finger gestures (Fig. 3). The word components are obtained with the help of a look-up table consisting of all inflectional

words available in the SIBI dictionary. The look-up table contains each inflectional word's affix and root word components [4]. In addition to the look-up table consisting of inflectional words, there is also a look-up table consisting of slang words, which will later be used to correct the input word.

Fig. 4 shows the implementation of the hand gesture text parser [15]. In-text parsing splits sentences into word components. This process starts with the text tokenizer, splitting sentences based on spaces between words. Next is miss-spelling correction using slang word table look-up. This process corrects the slang words into root words. Furthermore, after the sentence is split into word components and corrected, the Word Mapper process will use the look-up table created previously based on the SIBI dictionary to split any inflectional word into its affix and root word.

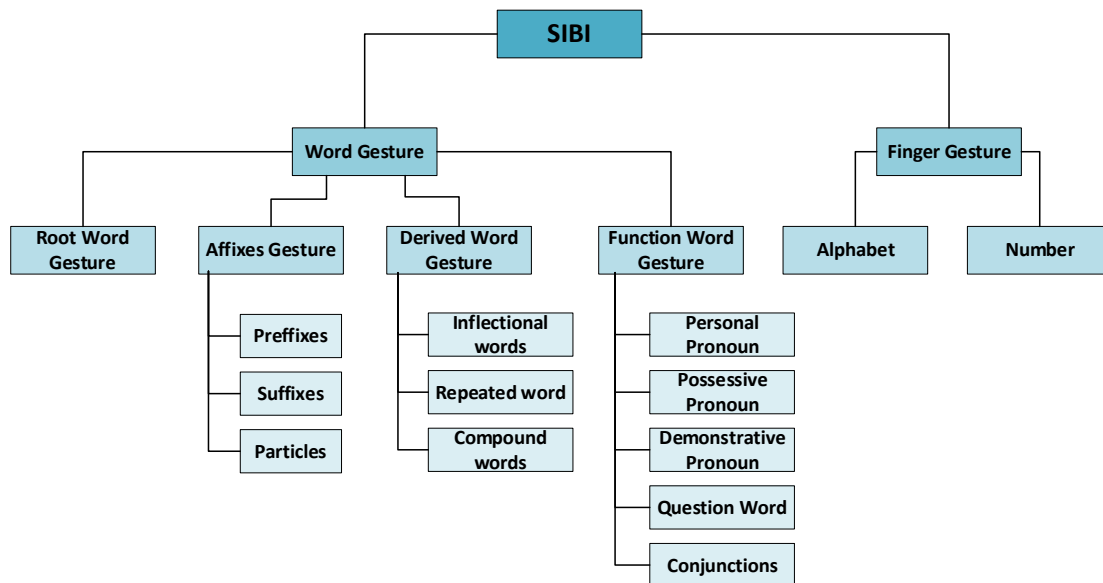


Fig. 3. Gestures in SIBI.



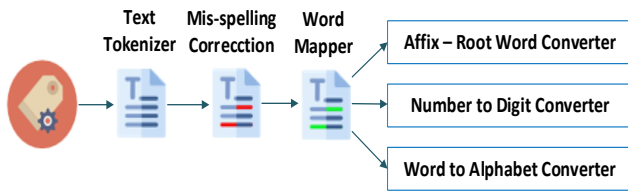


Fig. 4. Hand Gesture Text Parser Module.

The output of the hand gesture text parser, for example, is the splitting of the inflectional word "mengatakan" (= to say/ to tell) into prefix "me" + root word "kata" + suffix "kan." Meanwhile, the names and numbers in the sentence will be split into finger gestures. Names will be split according to the alphabet in the name; for example, William will be split into "w," "i," "l," "l," "i," "a," and "m." At the same time, the numbers will be split into the essential components of numbers, such as ten, hundred, thousand, and million: for example number "6203" will be split into "6" + "ribu" (= thousand) + "2" + "ratus" (= hundred) + "3" as seen in Fig. 4 [15].

2) *Mouth text parser*: The mouth text parser is a module used to break a sentence input into syllables. There are four steps to breaking the input sentence into syllables:

- Removing symbols and punctuation marks from the input sentence.
- Identifying the words and numbers in the sentence.
- Changing them to lowercase.
- Breaking the word into syllables according to the syllable look-up table (Table III below shows part of the syllable look-up table).

TABLE III. LIST OF INDONESIAN SYLLABLES AND THEIR EXAMPLES

Syllable	Example
V	a-tau, i-kan, u-ang, e-lang
CV	ba-ca, du-ka, ko-ta
VC	an-da, il-mu, ku-il, in-dah
CVC	tam-bah, sam-bal, tum-pah
CCV	pri-a, pu-tra pu-tri, tri-o
VCC	eks-tra, bu-ang
CVCC	teks,
CCVC	Stig-ma
CCCV	In-stru-men
CCCVC	Struk-tur
CCVCC	Kom-pleks, ke-nyang, me-nyang-kut

Note: C = consonant, V = vowel

Forming the mouth movements will use as many syllables as the syllables in the input sentence.

A syllable is a part of a word articulated in one breath. A syllable consists of one or more phonemes combined. Each syllable always contains a vowel phoneme [16][17]. Table III shows some examples of syllables. The actual syllable look-up table consists of all syllables in the SIBI dictionary.

In the previous study (Muzahidin and Rakun, 2020) [18], the respondents gave suggestions and criticisms of the lack of tongue movement. The tongue is essential in pronouncing a word [19]. Therefore, this study improves the mouth movement animation by adding tongue movement. With the addition of tongue movements, respondents can better distinguish words with similar lip movements. Tongue movements are formed separately from the formation of lip movements. The lip movement is formed based on videos of SIBI experts pronouncing words. On the other hand, the tongue movement data was formed according to the rules in the Bina Talk book, as shown in Table IV [20].

TABLE IV. TONGUE MOVEMENT

No	Name	Description	Example
1	Apiko dental	The tip of the tongue at the base of the upper teeth touches the front alveolar (upper gum).	/t/, /d/, /n/
2	Apiko alveolar	The tip of the tongue meets the arch of the tooth (alveolar)	/s/ and /z/
3	Dorso velar	Attach the back of the tongue to the area (soft palate)	/ng/, /g/, /k/, and /x/
4	Fronto platal	The center of the tongue is the articulator, and the palate is the articulation	/j/, /c/, and /y/
5	Lateral	Lifting the tongue to the palate	/l/
6	Vibration	Attaching the tongue to the alveolar (gums) and so on repeatedly	/r/

3) *Module text parser results*: The hand gesture and mouth movement text parser will produce different sentence fragments. Table V shows examples of output from hand gestures and mouth movement text parsers. Next, the output of each text parser will be used to generate hand gesture and mouth movement animations.

TABLE V. TEXT PARSER RESULT

Word	Hand Gesture	Mouth Movement	Represent
Saya	Saya	Sa, ya	Root Word
Surya	S, u, r, y, a	S, u, r, y, a	Fingerspelling (name)
45	Empat, puluh, lima	Em, pat, pu, luh, li, ma	Fingerspelling (number)
Abu-abu	Abu-abu	A, bu, a, bu	Repeated Word
Memakai	Me-, pakai	Me, pa, kai	Prefix + root word
Pakaian	Pakai, -an	Pa, kai, an	Root word + suffix
Memakaikan	Me-, pakai, -kan	Me, pa, kai, an	Prefix + root word + suffix

### C. 3D Animation Generation

This section discusses how to generate hand gesture animation based on a hand gesture text parser (1) and how to generate mouth movement animation based on a mouth movement text parser (2).

1) *Hand gesture animation engine*: Hand gesture data creation begins with hand movement data collection using

sensors from Perception Neuron v2. This study used 25 sensors: 3 upper-body, 1 left-shoulder, 3 left-hand, 7 right-fingers, 1 right-shoulder, 3 right-hand, 7 right-fingers. These sensors attached to a SIBI expert model will record data in coordinates of the position and rotation of the human body joints. All 3,100 words available in the SIBI dictionary were recorded. The recorded sensor data are then stored as skeletal animation. The data generated by the sensor is not perfect. Fig. 5 shows the difference between the hand movements made by the SIBI expert (right image) and the hand gestures generated by the sensor (left image). The AxisNeuron application is used to check and correct every skeletal data generated. In this process, the skeletal animation clip needs to be readjusted with references to the SIBI dictionary. The corrected skeletal animation clips were exported in fbx format and are used by Unity3D to generate hand gesture animation.



Fig. 5. Data Sensor Recording Result.

2) *Mouth animation engine*: The mouth animation movement was created based on SIBI expert facial data detected using the OpenPose<sup>12</sup> library. Then, the Dirichlet Free-Form Deformation (DFFD) uses the coordinate data from the face detection to build mouth movements. The data used in this process is data for mouth movements generation only. In addition, it is also necessary to develop tongue movements based on the Bina Bicara book.

a) *Mouth Movement*: The process of generating mouth movements starts by recording the SIBI expert pronouncing every word from the SIBI dictionary. Using actual video data can produce more realistic lip movements [21]. The steps to obtain the face coordinates to drive the three-dimensional mouth movement animation are as follows (as shown in Fig. 6 and Fig. 7):

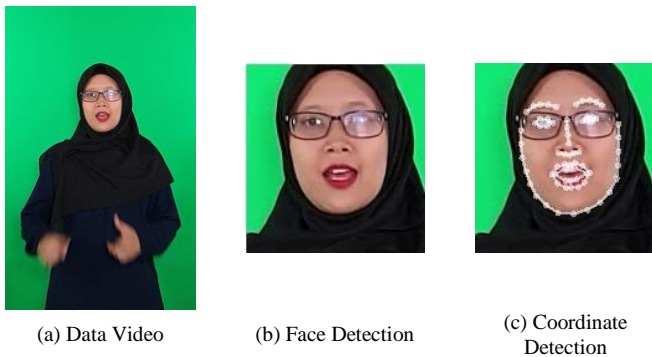
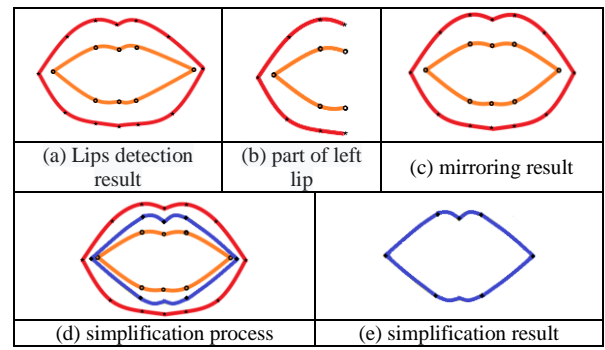


Fig. 6. Obtaining Face Coordinate Steps.



★/ red line = outer coordinates ( $po$ ); ●/ orange line = inner coordinates ( $pi$ ); ◆/ Blue line = middle coordinates ( $pm$ )

Fig. 7. Simplification Lip Point.

- Take sign language gesture with its pronunciation video performed by a SIBI expert [Fig. 6(a)].
- Crop the video to focus on the face only [Fig. 6(b)].
- Implement the coordinate detection process on the face using OpenPose [Fig. 6(c)].
- Get coordinates lip point [Fig. 7(a)].
- Cut the lips in half, and take a left part [Fig. 7(b)].
- Discard the right half of the lip and replace it with the mirror of the left lip to form symmetrical lips [Fig. 7(c)].
- Do the lips simplification process by averaging the outer and inner lips [Fig. 7(d)].
- The result of the lips simplification will be used for mouth movement animation [Fig. 7(e)].

The mirroring process is implemented because the shape of the human lips is not symmetrical between the left and right parts, caused by various factors such as teeth, cheeks, and face shape. The lip coordinates are then identified as the outer coordinates ( $po$ ) and the inner coordinates ( $pi$ ). Then, find the middle coordinates ( $pm$ ) using equation 1 [22]. Fig. 7(c) shows the results of the lip point simplification. These lip simplification coordinates will generate lip movement animation [18].

$$pm_x = \frac{(po_x + pi_x)}{2} \quad (1)$$

b) *Dirichlet Free-Form Deformation (DFFD)*: In this study, the Dirichlet Free-Form Deformation (DFFD) method is used to deform the 3D model. The application of DFFD in this research is by the following process:

- Apply the point of mouth movement form [Fig. 7(e)] into the 3D animation.
- Using the DFFD method to make mouth movements.
- Do this process for all the syllables in the SIBI dictionary
- Save this mouth movement into the database

<sup>1</sup><https://github.com/CMU-Perceptual-Computing-Lab/openpose>

The DFFD process changes the coordinates of a control point in an area that affects changes around another control point. Fig. 8 is the result of the deformation of the DFFD where the change in coordinates at one point has a local deformation area that reaches its surrounding points. All coordinate points move each other towards the specified coordinates. The results of these coordinates create changes in the shape of the lips according to the input data.

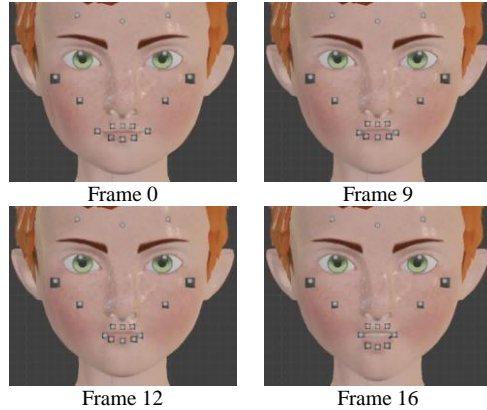


Fig. 8. 3D Animation Mouth Movement Syllable “Bu”.

Next is transferring the lip movement data into a 3D animation. Each lip part that will be moved has to be defined and coded for each syllable. All animation clips of Indonesian viseme syllables will be available after doing this process for all syllables. These visemes syllables animation will be stored as a database that the SIBI application can use.

The determining factor in understanding the word pronounced depends not only on the movement of the lips but also on teeth, tongue, and expression. Until now, no research has proven that people can catch a syllable or a word by looking at lip movements alone. Providing the 3D models with teeth and the tongue movement corresponding to each syllable will make it easier for the deaf to catch each syllable spoken.

#### D. Synchronizing Hand Gestures and Mouth Movement

Hand gestures and mouth movements are called based on the input sentence. The input sentence is deconstructed into words and syllables to generate hand gestures and mouth movements. The hand gestures and mouth movements will be generated simultaneously, but they do not take the same amount of time. So to synchronize hand gestures and mouth movements, it is necessary to accelerate or decelerate the movement of hand gestures. The speed of hand movement will follow the speed of the mouth movements. Generally, humans need 1 to 3 seconds to pronounce a word (depending on the length of the spoken word). Usually, words with only two syllables will be pronounced in one second. The gestures in the hands will follow the speed of the word's pronunciation. A one-second (two syllables) video is converted into frames,

which is 30 frames per second or equal to 15 frames per syllable. So the length of the hand gesture is 15 frames multiplied by the number of syllables of the word input. The algorithm for synchronizing hand gestures and mouth movements can be seen in Algorithm 1 below. The 3D animation will be displayed in a full model with synchronized pronunciation and hand gestures.

---

#### Algorithm 1 Synchronizing Hand and Mouth Movement

---

```

program start
initialize variable word_input
initialize variable mouth
initialize variable handsign
initialize variable frame = 15
start
    call function splittosyllable with word_input
    splittosyllable return value mouth
    output mouth

    handsign = mouth*frame
    output handsign

    call function runanimationmouth withinput mouth
    call function runanimationhandsign withinput
handsign
end
    
```

---

#### E. Insertion of Transition Movement using Cross Fade

This section explains how to create animated hand gestures. Hand gesture data creation begins with collecting hand and fingers movement coordinates for each word in the SIBI dictionary using a sensor, Perception Neuron v2. The coordinates of the hand and fingers are stored in a database. Creating hand gestures from an input sentence is done by retrieving the hand and fingers coordinates of each word in the sentence from the gesture database. Next is to insert transition gestures between words to create smooth, unified 3D animation sentence gestures. Fig. 9 shows the position of transition gestures in a sentence. This research uses the cross-fade method implemented using the Animancer API [15][18] to create transition gestures.

This research uses interpolation to generate a smoother transition movement between word gestures. Linear interpolation, also known as LERP, is the method to interpolate the positions and rotation values from the last frame of a word animation to the first frame of the following word animation linearly. It has the advantages of easy implementation and short execution time as the animation method traditionally used in animation [23]. Linear interpolation is a parametric curve defined as a straight line function can be seen in the following equation 2:

$$Q(u) = P_0 + u(P_1 - P_0) \tag{2}$$

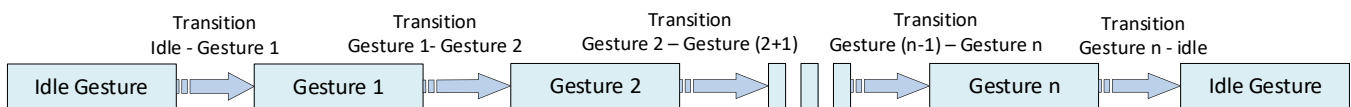


Fig. 9. Transition Motion between Word Gestures.

Equation (2) can also be written as:

$$Q(u) = (1 - u)P_0 + uP_1 \quad (3)$$

The value of  $u$  is used to set the interpolation to be built. If the value of  $u$  is 0, then  $Q(u)$  will be equal to the starting point of  $P_0$ , whereas if the value of  $u$  is 1, then  $Q(u)$  will be equal to the endpoint of  $P_1$ . If the value of  $u$  is between 0 and 1, then it will produce a point on the line  $\overline{P_0P_1}$ . Interpolation occurs if the value of  $u$  is in the interval  $[0,1]$ . If the value of  $u$  is outside the interval, it will not be interpolated.

Cross-fading is used to combine two LERP-based animated clips [24]. This technique can generate a smooth transition between animation clips. Cross-fade works by stacking the timeline of two animated clips, as seen in Fig. 10. The merging process requires a blend percentage  $\beta$  of the clips to be merged.  $\beta$  starts at 0 at time  $t_{start}$ . The meeting time between clip A and clip B is when a cross-fade process occurs. The value of  $\beta$  is then slowly incremented to 1 until time  $t_{end}$ . At that time, only the animation of clip B will appear. The time interval when the cross-fade is in progress is called the *blend time* ( $\Delta t_{blend} = t_{end} - t_{start}$ ).

#### F. Evaluation Metric

There are three tests carried out to measure the performance of the system being built. The first test is intended to measure the mouth movement animation when using syllables. The second test is carried out to measure the usability of this system. The third test is to measure the execution time.

1) *Evaluation of mouth movement*: Evaluation of mouth movement is done by calculating the Mean Opinion Score (MOS). Four online questionnaires (done during the Covid-19 pandemic locked down) need to be filled in by the respondents to check how well the 3D animation works: In Questionnaire 1, the animation pronounces 40 Indonesian words available in the SIBI dictionary; In Questionnaire 2, the animation pronounces 25 sentences long sentences; Questionnaire 3 compares the 3D animation with the original videos; In questionnaire 4, combine the mouth movement with hand gestures simultaneously. MOS respondents consisted of three deaf students and three SIBI teachers from the School for special needs SLB Santi Rama.

The evaluation uses subjective values from the teachers and deaf students because there is still no ground truth to test the correctness of lip movements. This application is intended for the Deaf, so it requires a direct assessment of the intended target. The MOS assessment uses a scale from 1 to 5 [2].

Calculations using MOS can be seen in the following equation 4:

$$MOS = \sum_{i=1}^N \frac{x(i).k}{N} \quad (4)$$

Where:

$x(i)$  = sample number  $i$

$k$  = weight value

$N$  = Number of Respondents

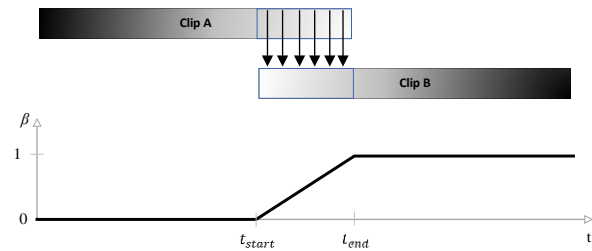


Fig. 10. Smooth Transition on Cross-Fading.

2) *Application survey measurement*: The System Usability Scale (SUS) test, a qualitative research tool, is used to assess and improve the usability of this system. Usability testing is carried out using a usability testing sheet containing tasks and scenarios the respondent must do during the test. In each task and scenario, the respondent will be assessed on whether he/she has succeeded in carrying out the task, the ease and difficulty the respondent faced, things the respondent likes and dislikes, and any suggestions from the respondent [25]. In each question, respondents will be asked to determine their rating on a scale of 1-5 based on their experience after using an interactive system design. A low score indicates disagreement from the respondent, while a high score indicates the respondent's agreement with the questions. SUS Score Assessment using the scoring formula [26]. Table VI is a list of questions of SUS taken from statements from [26] that have been translated into Indonesian [27]. The SUS itself consists of 10 items, the odd numbers are for positive items and the even numbers for negative. For positive items, the score contribution is the scale position minus 1 and for the negative items, the score contribution is 5 minus the scale position. The overall SUS score is the result of the sum of item score contributions multiply by 2.5, range from 0 to 100.



TABLE VI. TEXT PARSER RESULT

No	Question (Indonesian)	Question (English)
1.	Saya berfikir akan menggunakan sistem ini lagi	I think that I will use this system again
2.	Saya merasa sistem ini rumit untuk digunakan	I found the system to be unnecessarily complex
3.	Saya merasa sistem ini mudah untuk digunakan	I found the system easy to use
4.	Saya membutuhkan bantuan dari orang lain atau teknisi dalam menggunakan sistem ini	I think that I would need the support of a technical person to be able to use this system
5.	Saya merasa fitur-fitur sistem ini berjalan dengan semestinya	I found the various features in this system were well integrated
6.	Saya merasa ada banyak hal yang tidak konsisten(tidak serasi) pada sistem ini	I thought there was too much inconsistency in this system
7.	Saya merasa orang lain akan memahami cara menggunakan sistem ini dengan cepat	I feel that most people would learn to use this system very quickly
8.	Saya merasa sistem ini membingungkan	I found the system very cumbersome to use
9.	Saya merasa tidak ada hambatan dalam menggunakan sistem ini	I found no obstacles in the usage of this system
10.	Saya perlu membiasakan diri terlebih dahulu sebelum menggunakan sistem ini	I needed to learn a lot of things before I could get going with this system

#### IV. EXPERIMENT RESULTS

##### A. Mouth Movement Development

This research used deconstruction of words into syllables to generate mouth movements. Each syllable is developed based on lip movements and stored in a database. The lip movement of 3D animation starts from a silent mouth position labeled "idle." Then the lip movements are generated sequentially according to the word's syllables. At the end of the movement of the syllable, the mouth returns to the "idle" position. Fig. 11 shows an example of mouth movement generation. Furthermore, a transition motion that connects one syllable to the next using the cross-fading technique was added. This cross-fading technique is the same technique when generating transitions in hand movements.

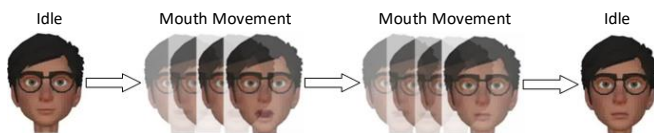


Fig. 11. Mouth Movement Generation.

##### B. Text-to-Gesture Application System Analysis

The text-to-gesture application system analysis measures the mean opinion score (MOS) on four questionnaires, System Usability Scale (SUS), and execution time. The MOS testing is used to test how well the mouth movement is in pronouncing words. Usability testing is a test to assess the user interface of the text-to-gesture application. Furthermore, execution time tests the time it takes to run the text-to-gesture application.

1) *Mean opinion score result:* The qualitative testing to measure the performance of the mouth movements generation on syllables was done by distributing four questionnaires to three SIBI teachers and three students from the School for special needs students, SLB Santi Rama. Questionnaire 1, that test the mouth movements animation in word pronunciation,

yields a score of 4.025. Questionnaire 2, which tests the mouth movements animation to pronounce a sentence, got a score of 4.025. Then, questionnaire 3, to test user understanding of the animation when hand gestures and mouth movements were combined, got a score of 4.422. Finally, questionnaire 4 tested the similarity between the animation and the original video and got a score of 4.282.

From the MOS results, respondents found it easier to catch words spoken solely (4.025) than in sentences (3.96). The combination of hand gesture and mouth movement animation can improve animation realism and understanding of the gestures demonstrated in 3D animation (3.96 vs. 4.422).

2) *System usability scale (SUS):* Google Form application is used to help design the questionnaire for SUS. The respondents will answer ten questions by choosing a scale from 1 – 5 for each item. This questionnaire was distributed for seven days and obtained 72 respondents. These respondents consisted of various backgrounds (sign language teachers, deaf people, and ordinary people), ranging from 19 to 61 years, and comprised 28 women and 44 men. The results of the SUS test obtained a score of 76.25 which means that it is categorized as GOOD and considered an application that users generally can accept [26].

3) *Execution time:* In building the Text-to-3D animation application, three processes involve the use of data, namely (1) sentence translation, (2) storage of application settings, and (3) dictionary search. In each of these processes, the execution time was tested 100 times. Furthermore, from the 100 results, a 95% confidence interval was calculated to find the actual execution time value interval. The 95% confidence interval results for each process successfully met the requirements to react instantaneously based on [23], which has an execution time of under 100 ms. The results of the execution time test can be seen in Table VII.

TABLE VII. EXECUTION TIME

Process	Confidence interval 95% execution time	Qualified react instantaneously (< 100 ms) based on research by Nielsen (1993)
Sentence translation	36.5 ± 2.65 ms	Yes
Storage of application settings	0.23 ± 0.039 ms	Yes
Dictionary search	2.21 ± 0.196 ms	Yes

## V. CONCLUSION

This study aims to build a SIBI Text-to-3D animation translator application system. The output of this application is SIBI 3D gesture animation of every input word in an Indonesian sentence. The 3D animation consists of hand gestures and mouth movements' animation. Hand gesture data creation begins with hand movement data collection using sensors from Perception Neuron v2. The recorded sensor data are then stored as skeletal animation. The mouth animation movement was created based on SIBI expert facial data detected using the OpenPose library. Then, the Dirichlet Free-Form Deformation (DFFD) uses the coordinate data from face detection to build mouth movements. Hand gestures and mouth movements are called based on the input sentence. The input sentence is deconstructed into words to generate hand gestures and syllables to generate mouth movements.

It is necessary to accelerate or decelerate the movement of hand gestures to synchronize hand gestures and mouth movements. This research uses Cross-Fade interpolation to generate a smoother transition movement between word gestures. There are three tests carried out to measure the performance of the system being built. The first test is intended to measure the mouth movement animation when using syllables by calculating MOS. The second test is carried out to measure system's usability by using the SUS test. The third test measures the execution time by calculating the time needed by processes involving data. From the MOS results, respondents found it easier to catch words spoken solely (4.025) than in sentences (3.96). The combination of hand gestures and mouth movement animation can improve animation realism and understanding of the gestures demonstrated in 3D animation (3.96 vs. 4.422). The resulting animation is quite similar to the original video (4.282). SUS score is 76.25, which means this application is in the GOOD category. The execution time of all processes that involved data (sentence translation, storage of application settings, and dictionary search) are less than 100ms, which means it met the application requirements to react instantaneously. Hopefully, this application can be used to solve the communication problems between the deaf and the people around them. Because the number of words available in the SIBI dictionary (around 3100 words) is far less than the Indonesian words, some words need to be fingerspelled or replaced by similar words available in the SIBI dictionary. In the future, a synonym table can be added to this application. The synonym table will speed up the process of replacing a word with its synonym from the SIBI dictionary. This synonym table must be updated regularly to cover as many Indonesian words as

possible. Another thing that can be done to improve this application is to add facial expressions to the 3D model. In sign language, facial expressions are used to strengthen the meaning of a sentence, just like intonation in spoken language.

## ACKNOWLEDGMENT

This work is supported by The National Research and Innovation Agency (BRIN) Implementation Grant, Number PKS-175/UN2.INV/HKP.05/2021. This support is gratefully received and acknowledged.

## REFERENCES

- [1] V. Khetani, Y. Gandhi, and R. R. Patil, "A Study on Different Sign Language Recognition Techniques," in 2021 International Conference on Computing, Communication and Green Engineering (CCGE), 2021, pp. 1–4.
- [2] S. Sumpeno, M. Hariadi, and A. M. Syarif, "Development of Indonesian Text-to-Audiovisual Synthesis System Using Syllable Concatenation Approach to Support Indonesian Learning," pp. 166–184, 2017.
- [3] K. Anggraini, E. Rakun, and L. Y. Stefanus, "Recognizing The Components of Inflectional Word Gestures in Indonesian Sign System known as SIBI (Sistem Isyarat Bahasa Indonesia) by using Lip Motion," in 2019 International Conference on Electrical Engineering and Informatics (ICEEI), 2019, pp. 384–389.
- [4] M. Adriani, J. Asian, B. Nazief, S. M. M. Tahaghoghi, and H. Williams, "Stemming Indonesian: A Confix-Stripping Approach," ACM Trans. Asian Lang. Inf. Process., vol. 6, 2007.
- [5] K. Karpouzis, G. Caridakis, S.-E. Fotinea, and E. Efthimiou, "Educational Resources and Implementation of a Greek Sign Language Synthesis Architecture," Comput. Educ., vol. 49, no. 1, pp. 54–74, Aug. 2007.
- [6] M. Boulares and M. Jemni, "Mobile Sign Language Translation System For Deaf Community 1–4.," 2012.
- [7] E. Efthimiou et al., "Sign Language Recognition, Generation, and Modelling: A Research Effort with Applications in Deaf Communication," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 5614 LNCS, no. PART 1, pp. 21–30, 2009.
- [8] S. Siswomartono, "Cara Mudah Belajar SIBI (Sistem Isyarat Bahasa Indonesia)," 2007.
- [9] A. Thangthai, B. Milner, and S. Taylor, "Synthesising Visual Speech Using Dynamic Visemes and Deep Learning Architectures," Comput. Speech Lang., vol. 55, pp. 101–119, 2019.
- [10] E. Setyati, O. Susandono, L. Zaman, Y. M. Pranoto, S. Sumpeno, and M. H. Purnomo, "Establishment of Indonesian Viseme Sequences Using Hidden Markov Model Based on Affection," 2017 Int. Semin. Intell. Technol. Its Appl. Strength. Link Between Univ. Res. Ind. to Support ASEAN Energy Sect. ISITIA 2017 - Proceeding, vol. 2017-Janua, pp. 275–280, 2017.
- [11] H. Haryanto and S. Sumpeno, "A Realistic Visual Speech Synthesis for Indonesian Using a Combination of Morphing Viseme and Syllable Concatenation Approach to Support Pronunciation Learning," vol. 13, no. 8, pp. 19–37, 2018.
- [12] J. Yu and Z. F. Wang, "A Video, Text, and Speech-Driven Realistic 3-D Virtual Head for Human-Machine Interface," IEEE Trans. Cybern., vol. 45, no. 5, pp. 977–988, 2015.
- [13] H. Ni and J. Liu, "3D Face Dynamic Expression Synthesis System Based on DFFD," in 2019 IEEE 3rd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC), 2019, pp. 1125–1128.
- [14] S. Taylor et al., "A Deep Learning Approach for Generalized Speech Animation," ACM Trans. Graph., vol. 36, no. 4, 2017.
- [15] E. Rakun and I. S. A. Darmana, "Generating of SIBI Animated Gestures from Indonesian Text," PervasiveHealth Pervasive Comput. Technol. Healthc., pp. 256–264, 2020.



- [16] S. Suyanto, "Incorporating Syllabification Points into a Model of Grapheme-to-Phoneme Conversion," *Int. J. Speech Technol.*, vol. 22, no. 2, pp. 459–470, 2019.
- [17] S. Suyanto, "Flipping Onsets to Enhance Syllabification," *Int. J. Speech Technol.*, vol. 22, no. 4, pp. 1031–1038, 2019.
- [18] S. Muzahidin and E. Rakun, "Text-Driven Talking Head Using Dynamic Viseme and DFFD for SIBI," in *2020 7th International Conference on Information Technology, Computer, and Electrical Engineering (ICITACEE)*, 2020, pp. 173–178.
- [19] L. Zhao and L. Czap, "Visemes of Chinese Shaanxi xi'an Dialect Talking Head," *Acta Polytech. Hungarica*, vol. 16, no. 5, pp. 173–193, 2019.
- [20] E. Sadjah, Bina Bicara, *Persepsi Bunyi dan Irama*. Bandung: PT. Refika Aditama, 2013.
- [21] I. R. Ali, H. Kolivand, and M. H. Alkawaz, "Lip Syncing Method for Realistic Expressive 3D Face Model," *Multimed. Tools Appl.*, vol. 77, no. 5, pp. 5323–5366, 2018.
- [22] M. Liyanthy, H. Nugroho, and W. Maharani, "Realistic Facial Animation of Speech Synchronization for Indonesian Language," *2015 3rd Int. Conf. Inf. Commun. Technol. ICoICT 2015*, pp. 563–567, 2015.
- [23] S. Lim, "Linear Interpolation Transition of Character Animation for Immediate 3D Response to User Motion," *Int. J. Contents*, vol. 11, no. 1, pp. 15–20, 2015.
- [24] J. Gregory, *Game Engine Architecture*, 3rd ed. CRC Press, 2019.
- [25] A. K. Darmawan, M. A. Hamzah, B. Bakir, M. Walid, A. Anwari, and I. Santosa, "Exploring Usability Dimension of Smart Regency Service with Indonesian Adaptation of The System Usability Scale (SUS) and User Experience Questionnaire (UEQ)," in *2021 International Conference on Computer Science, Information Technology, and Electrical Engineering (ICOMITEE)*, 2021, pp. 74–79.
- [26] Nielsen, *Usability Engineering*. San Diego, 1993.
- [27] Z. Sharfina and H. B. Santoso, "An Indonesian adaptation of the System Usability Scale (SUS)," in *2016 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, 2016, pp. 145–148.
- [28] Y. Bouzid and M. Jemni, "TuniSigner: A Virtual Interpreter to Learn sign Writing," *Proc. - IEEE 14th Int. Conf. Adv. Learn. Technol. ICALT 2014*, pp. 601–605, 2014.

# A Review of Foreground Segmentation based on Convolutional Neural Networks

Pavan Kumar Tadiparthi<sup>1</sup>, Sagarika Bugatha<sup>2</sup>, Pradeep Kumar Bheemavarapu<sup>3</sup>

Associate Professor, Department of Information Technology, MVGR College of Engineering (A), Vizianagaram, A.P, India<sup>1</sup>

Student, Department of Information Technology, MVGR College of Engineering (A), Vizianagaram, A.P, India<sup>2,3</sup>

**Abstract**—Foreground segmentation in dynamic videos is a challenging task for many researchers. Many researchers worked on various methods that were traditionally developed; however, the performance of those state-of-art procedures has not yielded encouraging results. Hence, to obtain efficient results, a deep learning-based neural network model is proposed in this paper. The proposed methodology is based on Convolutional Neural Network (CNN) model incorporated with Visual Geometry Group (VGG) 16 architecture, which is further divided into two sections, namely, Convolutional Neural Network section for feature extraction and Transposed Convolutional Neural Network (TCNN) section for un-sampling feature maps. Then the thresholding technique is employed for effective segmentation of foreground from background in images. The Change Detection (CDNET) 2014 benchmark dataset is used for the experimentation. It consists of 11 categories, and each category contains four to six videos. The baseline, camera jitter, dynamic background, and bad weather are the categories considered for the experimentation. The performance of the proposed model is compared with the state-of-the-art techniques, such as Gaussian Mixture Model (GMM) and Visual Background Extractor (VIBE) for its efficiency in segmenting foreground images.

**Keywords**—Foreground segmentation; deep learning; Convolutional Neural Network (CNN); Visual Geometry Group (VGG) 16 architecture; Transposed Convolutional Neural Network (TCNN); Gaussian Mixture Model (GMM); Visual Background Extractor (VIBE)

## I. INTRODUCTION

Foreground segmentation [1] is a major part of various applications of computer vision. Foreground segmentation means that segmenting moving information from static information (background). Foreground segmentation is also called as Background Subtraction or Change Detection. Foreground segmentation is widely used in several applications like video surveillance [2], traffic monitoring, shopping malls, airports, etc. It analyzes a video sequence by using a set of techniques and those video sequences are recorded by a stationary camera.

Gaussian Mixture Model introduced by Chris Stauffer et al., [3] works on pixel-based classification. GMM models each pixel with K-Gaussians. It easily copes up with illumination changes. Even though a single Gaussian function is not able to deal with a dynamic background by providing a low updating rate of the background model. It is failed by the camouflage effect. The number of Gaussians here is predetermined as either 3, 4 or 5.

Visual Background Extractor introduced by O. Barnich et al., [4] is a non-parametric method and it works on pixels. This method utilizes the spatial information around the pixel for the background model. First of all, a set of values should be taken for each pixel at the same location in the neighborhood. Later, it compares this set to the current pixel value to determine whether it is background or foreground and adapts the model by choosing randomly among values to substitute from the background model. This approach differs from classical approach and belief that the oldest values should be replaced first. Finally, when the pixel is found to be part of the background, its value is generated into the background model of a neighboring pixel. Visual Background Extractor (VIBE) applied to color values of pixels of background training sequences as samples of observed backgrounds. Visual Background Extractor (VIBE) shows the best performance because it using samples as background models to represent the background changes. However, Visual Background Extractor (VIBE) has a major disadvantage that it uses color values of pixels to build the background model but color values are found to be sensitive to noise and illumination changes.

This article concentrated on developing a foreground segmentation model based on Deep Learning technique called Convolutional Neural Networks. The Convolutional Neural Network (CNN) is associated with Transposed Convolutional Neural Network (TCNN) for extracting the feature maps to identify foreground image. A thresholding technique is utilized to filter out the feature maps which distinguishes foreground object from a background object in an image. The main contributions of this article are:

- Detecting foreground objects in an image with improved accuracy.
- Design and implementation of a model using Deep Learning technique for effective segmentation of foreground objects.
- Evaluation of the proposed model using Gaussian Mixture Model (GMM) & Visual Background Extractor (VIBE) techniques.

The reminder of the paper is further organized as follows: Section II discusses related work, Section III presents the proposed methodology, Section IV of the article illustrated the experimental setup, and Section V illustrates the performance evaluation and experimentation results. Section VI gives out the conclusion and future work.

## II. RELATED WORK

In the past few years, eminent researchers experimented on foreground object segmentation techniques such as, Midhula Vijayan et al., [5] proposed a deep-neural network architecture using temporal and spatial information from background images and current processing images. Their experimentation obtained a better performance model when compared with existing background subtraction methods both qualitatively and quantitatively. Tsung-Han Tsai et al., [6] elucidated an unsupervised segmentation technique using distinct thresholding techniques by sorting of changed pixels ratio for precise decision. They obtained satisfactory results in generic images. Patrick Dickinson et al., [7] addressed the problem of foreground segmentation, when there is a varying background over time using Adaptive Gaussian Mixtures model (AGMM). Their experimentation resulted in better performance than the per-pixel and Markov Random Field-based Models and achieved a Jaccard coefficient of 0.59.

Jaime Gallego et al., [8] experimented by combining pixel-wise and region-based model for foreground segmentation using one Gaussian per pixel. This improved the performance of the system having similar colors to those of the background. Jaime Gallego et al., [9] proposed a method for monocular static camera sequences and indoor scenarios using Gaussian pixel color, Modified Mean Shift algorithm, and by Bayesian framework. Their methodology yielded robust segmentation and tracking of objects than the state-of-art methodologies. Jaime Gallego et al., [10] illustrated the reduction of false positive and false negative by using region-based models for modeling foreground and background region. Their model surrounds the foreground by Maximum A Posteriori and Markov Random Field model which uses pixel-wise color GMM for background sequence classification.

Jiayu Liang et al., [11] constructed a new method using Genetic Programming for feature construction by incorporating subtree technique. The simultaneous construction of multiple features and parsimony pressure techniques are introduced to improve the proposed techniques bloat control. Xuchao Gong et al., [12] developed a method using the GMM for modeling static background regions and inter-frame change detection and Scale-Invariant Feature Transform (SIFT) feature analysis is used for boundary identification of foreground regions. The Grab cut methods are used for segmentation of foreground moving objects. Yizheng Guo et al., [13] illustrated the segmentation of pigs in group-housed environments by combining a mixture of Gaussians using prediction mechanism and threshold segmentation algorithm.

Nikolaos Katsarakis et al., [14] considered Stauffer and Grimson's algorithm as a baseline algorithm and enhanced their algorithm by changing the learning rate and combining the Gaussian mixture if they are similar. They yielded good results than the baseline algorithm.

As per the literature review, many models proposed by eminent researchers have failed to obtain efficient results for foreground segmentation. This paper focuses on the implementation of foreground segmentation methodology using deep learning techniques for enhanced segmentation results.

## III. PROPOSED METHODOLOGY

The proposed methodology uses the VGG-16 model for foreground segmentation. It comprises two sections namely Convolutional Neural Network (CNN) and a Transposed Convolutional Neural Network (TCNN).

Each frame is extracted from video data and moved onto the following Convolutional Neural Network (CNN) and Transposed Convolutional Neural Network (TCNN).TCNN networks, where, the Convolutional Neural Network (CNN) network extracts the features from the input frame which are related to the foreground object by moving the frame onto different layers which are described in section A. The extracted features from the Convolutional Neural Network (CNN) network are fed into Transposed Convolutional Neural Network (TCNN), where the feature maps are unsampled to obtain original input size. By applying the thresholding technique based on probability to the unsampled features, the foreground objects are segmented from the background objects in a frame and hence for all the frames.

The proposed model is highlighted in Fig. 1. The Convolutional Neural Network and Transposed Convolutional Neural Network are described as follows:

### A. Convolutional Neural Network (CNN)

The actual VGG-16 model contains 5 blocks with 16 layers. The feature extractor section uses 5 blocks and each block contains a set of  $3 \times 3$  kernels like a stack with a max-pooling layer. The filters used in each convolutional layer are 64, 128, 256, 512, and 1024.

Our proposed model contains only 4 blocks as a feature extractor section and each block having  $3 \times 3$  kernels as a smaller kernel size. Two receptive fields of  $3 \times 3$  kernels are equal to a  $5 \times 5$  receptive field. The three  $3 \times 3$  kernels are equal to the  $7 \times 7$  receptive field. Due to this, the parameters are reduced to 30%. The input is a  $W \times H$  RGB image. The four convolutional layers of  $3 \times 3$  kernels are followed by a max-pooling layer with stride 2. This first section produces the feature maps of size  $W/8 \times H/8$  with 512. Doing this input by half, we get fine features from each frame, so it will have minute information from frame. The feature extraction section is done well because the Visual Geometry Group (VGG) 16 architecture, network is pre-trained. It is already trained on millions of images. It uses the CDNET-2014 dataset as input. The obtained feature maps are fed as input for the Transposed Convolution Neural Network (TCNN) part.

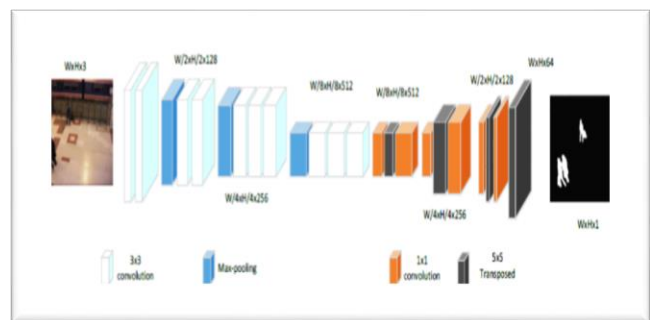


Fig. 1. Architecture of the Experimented CNN.

### B. Transposed Convolutional Neural Network (TCNN)

The feature maps in the CNN section will be unsampled by Transposed Convolution Neural network (TCNN). It unsamples the features into original input by performing the multiplication of output by transposing the kernel or padding output to reconstruct the input. It down samples the feature maps by decreasing them from 512 to 64.

In this TCNN section, we have four blocks each block contains two 1x1 convolutions followed by 5x5 transposed convolutions with stride 2. These 1x1 convolutions are used to shrink the feature maps to get the original input size. After that thresholding technique is applied for the segment the foreground object.

### IV. EXPERIMENTAL SETUP

The overall experimentation was carried out on a 64-bit Windows 10 operating system having Inter® core™ i5 processor clocked at 2.24 GHz, 8 GB RAM, and 1 TB hard drive installed with Anaconda, Python platform with Tensor flow as backend and supporting image processing packages.

The CDNET-2014 benchmark dataset is used for experimentation. The dataset comprises of 11 categories and each category has 4 or 5 videos. For the experimentation, only four categories are namely: baseline, camera jitter, bad weather, and dynamic background to validate the performance of the model.

### V. PERFORMANCE EVALUATION AND EXPERIMENTATION RESULTS

To evaluate the performance of the proposed model different quality metrics are considered such as: Precision, Recall, Accuracy, F-Score, mean Squared Error (MSE), Root Mean Squared Error (RMSE), False Negative Rate (FNR), False Positive Rate (FPR), Peak Signal to Noise Ratio (PSNR), and Pair Wise Correlation (PWC) coefficient which are defined as follows:

$$Precision = \frac{True_{pos}}{False_{pos} + True_{pos}} \quad (1)$$

$$Recall = \frac{True_{pos}}{False_{neg} + True_{pos}} \quad (2)$$

$$Accuracy = \frac{True_{pos} + True_{neg}}{True_{pos} + False_{pos} + True_{neg} + False_{neg}} \quad (3)$$

$$F - Score = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (4)$$

$$MSE = \frac{1}{n} \sum_{i=1}^n (Actual\ Values - Predicted\ Values)^2 \quad (5)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (Actual\ Values - Predicted\ Values)^2} \quad (6)$$

$$FNR = \frac{False_{neg}}{True_{pos} + False_{neg}} \quad (7)$$

$$FPR = \frac{False_{pos}}{True_{neg} + False_{pos}} \quad (8)$$

$$PSNR = 10 \log_{10} \frac{R^2}{MSE} \quad (9)$$

$$PWC = 100 * \frac{False_{neg} + False_{pos}}{True_{pos} + False_{neg} + True_{neg} + False_{pos}} \quad (10)$$

To validate the performance of the proposed model, GMM and VIBE are considered as reference model. The graphical illustration of experimental results is shown in Fig. 2 and their corresponding results are given in Tables I to IV.















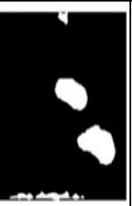





Dataset	Baseline	Bad Weather	Camera Jitter	Dynamic Background
	Highway	Skating	Boulevard	Boats
Input				
Ground Truth				
GMM				
VIBE				
CNN				

Fig. 2. Foreground Detection of Baseline, Camera Jitter, Dynamic Background and Bad Weather from CDNET-2014.

TABLE I. EVALUATION METRICS OF DIFFERENT METHODS ON CAMERA JITTER FROM CDNET DATASET

Metrics/Methods	GMM	VIBE	CNN
Precision	0.0127	0.0168	0.0197
Recall	0.124	0.0427	0.0314
Accuracy	0.9155	0.9853	0.9998
F-Score	0.0152	0.0372	0.0584
MSE	0.029	0.025	0.021
RMSE	0.0541	0.0502	0.0122
FPR	0.0845	0.046	0.025
FNR	0.743	0.852	0.975
PSNR	73.5035	74.1422	86.4039
PWC	6.4477	4.4724	2.0219

TABLE II. EVALUATION METRICS OF DIFFERENT METHODS ON DYNAMIC BACKGROUND FROM CDNET DATASET

Metrics/Methods	GMM	VIBE	CNN
Precision	0.0138	0.0154	0.0185
Recall	0.132	0.0316	0.0213
Accuracy	0.9725	0.9836	0.9999
F-Score	0.0131	0.0281	0.0673
MSE	0.039	0.036	0.034
RMSE	0.0430	0.0325	0.0132
FPR	0.0032	0.023	0.045
FNR	0.621	0.743	0.864
PSNR	71.824	72.9144	91.1751
PWC	5.3367	4.6399	3.0136

TABLE III. EVALUATION METRICS OF DIFFERENT METHODS ON BASELINE FROM CDNET DATASET

Metrics/Methods	GMM	VIBE	CNN
Precision	0.0154	0.0162	0.0174
Recall	0.135	0.0538	0.0125
Accuracy	0.9412	0.9861	0.9998
F-Score	0.0125	0.0473	0.0762
MSE	0.0204	0.047	0.053
RMSE	0.0571	0.0492	0.0100
FPR	0.0588	0.0137	0.077
FNR	0.632	0.835	0.952
PSNR	65.0579	74.3177	88.1648
PWC	5.8794	0.3922	0.0154

TABLE IV. EVALUATION METRICS OF DIFFERENT METHODS ON BAD WEATHER FROM CDNET DATASET

Metrics/Methods	GMM	VIBE	CNN
Precision	0.0134	0.0184	0.0195
Recall	0.128	0.0724	0.0512
Accuracy	0.9202	0.9674	0.9999
F-Score	0.0236	0.0584	0.0873
MSE	0.0408	0.0219	0.0132
RMSE	0.1442	0.0439	0.0133
FPR	0.0798	0.0526	0.0266
FNR	0.543	0.721	0.843
PSNR	64.9842	75.3205	93.4217
PWC	7.9846	4.2576	0.0143

## VI. CONCLUSION AND FUTURE WORK

This paper focused on the enhancement of foreground object segmentation using deep neural network model viz., VGG-16 which comprises of CNN and TCNN. To evaluate the performance of the proposed model, comparison with the traditional methods such as GMM and VIBE is done. The results showcased that the proposed VGG-16 model has proven its supremacy in obtaining the highest accuracy of 99.99% in - comparison to the state-of-the-art techniques such as GMM and VIBE. Hence, the proposed model performed better in segmenting the foreground image with improved accuracy. Exploring other advanced deep learning techniques for improved segmentation of foreground objects in images is considered as the future work.

### REFERENCES

- [1] Shahbaz, L. Kurniangggoro, Wahyono, and K.-H. Jo, "Recent Advances in the Field of Foreground Detection: An Overview," Advanced Topics in Intelligent Information and Database Systems. Springer International Publishing, (2017), pp. 261–269.
- [2] Shahbaz, J. Hariyono, and K. H. Jo, "Evaluation of background subtraction algorithms for video surveillance," 21st Korea-Japan Joint Workshop on Frontiers of Computer Vision (FCV), Jan (2015), pp. 1–4.
- [3] Chris Stauffer, W.E.L. Grimson, "Adaptive background mixture models for real-time tracking," Proceedings of IEEE Conference Computer vision Pattern Recognition, Vol. 2, Jan, (2007).
- [4] O. Barnich, M. Van Droogenbroeck, "ViBe: A universal background subtraction algorithm for video sequences," IEEE Transactions on Image Processing, Vol. 20: Issue. 6, Jun (2011).
- [5] Midhula Vijayan, R. Mohan, "A Universal Foreground Segmentation Technique using Deep-Neural Network", Multimedia Tools and Applications, May, (2020).
- [6] Tsung-Han Tsai, Guan-Jun Chen, Wen-Liang Tzeng, "A Novel Foreground/Background Decision using in Unsupervised Segmentation of Moving Objects in Video Sequences", 46th Midwest Symposium on Circuits and Systems, Cairo, Vol. 3, Dec, (2003).
- [7] Patrick Dickinson, Andrew Hunter, Kofi Appiah, "A spatially distributed model for foreground segmentation", Image and Vision Computing, Vol. 27: Issue. 9, Aug, (2009).
- [8] Jaime Gallego, Montse Pardo, Gloria Haro, "Bayesian Foreground Segmentation and Tracking using Pixel-wise Background Model and Region based Foreground model", 16th IEEE International Conference on Image Processing (ICIP), Cairo, Nov, (2009).
- [9] Jaime Gallego, Montse Pardo, Gloria Haro, "Enhanced foreground segmentation and tracking combining Bayesian background, shadow and foreground modeling", Pattern Recognition Letters, Vol. 33: Issue 12, Sep, (2012).
- [10] Jaime Gallego, Pascal Bertolino, "Foreground Object Segmentation for moving camera sequences based on foreground probabilistic models and prior probability maps," IEEE International Conference on Image Processing (ICIP), Paris, Oct, (2014).
- [11] Jiayu Liang, Yu Xue, Jianming Wang, "Genetic programming-based feature construction methods for foreground object segmentation," Engineering Applications of Artificial Intelligence, Vol. 89, Mar, (2020).
- [12] Xuchao Gong, Zongmin Li, "Efficient Foreground Segmentation Using an Image Matting Technology," 2013 International Conference on Computational and Information Sciences, Shiyang, Jun, (2013).
- [13] Yizheng Guo, Weixing Zhu, Pengpeng Jiao, Jiali Chen, "Foreground detection of group-housed pigs based on the combination of Mixture of Gaussians using prediction mechanisms and threshold segmentation," Biosystems Engineering, Vol. 125, Sep, (2014).
- [14] Nikolaos Katsarakis, Zheng-Hua Tan, Ramjee Prasad, Aristodemos Pnevmatikakis, "Improved Gaussian Mixture Models for Adaptive Foreground Segmentation," Wireless Pers Commun 87, Apr, (2016).



# Multi-instance Finger Knuckle Print Recognition based on Fusion of Local Features

Amine AMRAOUI<sup>1\*</sup>, Mounir AIT KERROUM<sup>2</sup>, Youssef FAKHRI<sup>3</sup>  
LaRI Laboratory, Faculty of Science, Ibn Tofail University, Kenitra, Morocco<sup>1,3</sup>  
LaRI Laboratory, ENCG, Ibn Tofail University, Kenitra, Morocco<sup>2</sup>

**Abstract**—Personal identity has become an important asset in today's digital world for any individual in society. Biometrics offers itself as a reliable and secure guarantor of our identities, so it has become essential to build efficient and robust recognition systems. In this orientation, we propose a fusion approach, which aims to optimally exploit the dividing block dimensions in the case of local methods to reduce similarities. We will use the compound local binary model (CLBP) for local features extraction, a robust operator descriptor that exploits both the sign and the inclination information of the differences between the center and the neighbor gray values. The reliability of the proposed approach was evaluated on the PolyU Finger Knuckle Print (FKP) database. We presented several experimental results that show the detailed path of our approach, explain the choices made for each step and illustrate the significant improvements compared to other existing recognition systems in the literature. The recognition rate of the proposed global approach is one of the highest among the other methods. Optimal final approach recognition rates vary between 99.70% and 100%.

**Keywords**—Biometrics; Finger Knuckle Print; local features; fusion; compound local binary pattern

## I. INTRODUCTION

The security of personal identity has become a major asset in the development of the world we live in. This era has a wide spectrum of daily transactions that are in the billions, given the number of the world's population involved in the digital world and its applications. This digital world is committed to the development of many services, the main purpose of which is to facilitate this mass of interaction between the population and the services, and to ensure the efficiency of all the transactions that can be judging sensitive; such as finance and communication. These needs are generally linked to many risks, particularly security. The implementation of mechanisms and efficient applications to ensure personal identity has become important and urgent given the daily risks. To overcome these risks, the use of biometrics is an effective way to solve security difficulties in various fields and services [1].

During the last decades, several research works have been conducted to build reliable recognition systems. Researchers have exploited and experimented with various biometric modalities including face, voice, fingerprint, palm print, iris, [2-6], etc. Some types of these biometric descriptors show an intrusive nature [7]. It should be noted that the acceptance and ease of use of biometric identifiers play a key role in the success of recognition systems. To ensure these two points, the thinking of the researchers was drawn to the hand-based

descriptors. There is a lot of research and promising results in the literature on hand-based biometric modalities, e.g. hand [8], [9], palm print [10], [11], fingerprint [12], [13], and hand geometry [14], [15], which have been widely studied. These studies have allowed the construction of a large set of recognition systems to ensure the identity authentication function and which operate successfully in several areas.

The most common identification systems in the real world are based on the use of fingerprint recognition, moreover it is the most used system in the field of access control, the police etc. Moreover, the most reliable systems are based on the use of the iris as an identification modality. Except that, these two descriptors represent drawbacks, which can hinder their success, iris sensors represent high intrusiveness, which makes them not very acceptable by users and extraction of small unique features called minutiae from damaged fingerprints is difficult [16]. It is added to the two drawbacks cited, the fact that fingerprints are vulnerable to spoofing attacks which consist in the creation of biometric artefacts. Matsumoto et al. [17] found that gummy fingers were accepted with high recognition rates by the 11 different fingerprint systems they used. These forged fingerprints are easily achievable with readily available devices and materials. This vulnerability is due to the anatomical structure of the hand and the mechanisms of its movements; in fact, the use requires a contact surface between the hand bottom and the object used, this interaction will keep traces of fingerprints and palm prints on this object. To overcome this problem, one of the proposed solutions is to employ the back side of the hand. In recent years, researchers have observed that the skin pattern of the outer surface of the fingers, especially in the area around the phalangeal joint, has a rich texture due to the lines and folds of the skin. This texture shows a distinctive character given the uniqueness it represents; therefore the finger knuckle print can be used as a biometric descriptor [18], [19].

In the literature, researchers have begun their work to create recognition systems based on finger knuckle print FKP, with approaches based on the use of global characteristics such as: principal component analysis (PCA), independent component analysis (ICA) and linear discriminant analysis (LDA) [20]. Subspace analysis approaches are methods whose concept is suitable for large devices; they are rather effective for large areas such as facial recognition, which reduces their performance for systems that use descriptors with smaller surfaces such as FKP images [21]. Subsequently, researchers turned to multi-algorithm or multimodal approaches [22-24]. The mechanisms based on these approaches have shown

\*Corresponding Author.



satisfactory results in the majority of cases. Based on this observation, we directed our work towards FKP recognition systems using local and not global characteristics, given the nature of the descriptor. Also, our orientation to ensure maximum reliability was to continue work towards the construction of a multi-instance biometric system, based on a reliable and robust extraction algorithm.

The response to these established expectations was the motivation we set ourselves to propose an efficient approach, which relies on the use of local characteristics to build a reliable multi-instance recognition system and which allows reducing the complexity of the calculations, as well as the implementation cost will be optimized. The proposed approach is based on the use of the compound local binary model (CLBP) [25] for the assurance of the local feature extraction phase. In this phase, we experimented and evaluated this method on several block sizes and also the nature of the block (square or rectangular), which is often neglected in other works, to produce optimal histograms and achieve the best possible results. The classification phase will be oriented towards classifiers based on the use of distances, this kind of classifiers can produce good results in the case of low resolution images, which we will use. We have opted for three variants of measurements: Euclidean distance, Jeffrey divergence and city block, which we will study their performance with our extracted characteristics and choose the most suitable to work in a real-time environment and produce effective results.

This paper is organized as follows: in Section 2, we will describe the proposed approach and the methods used. In Section 3, we will report and discuss the experimental results conducted on the PolyU FKP database [26] and finally, in Section 4, we will draw our conclusions.

## II. PROPOSED APPROACH

The construction of recognition systems in the real world is based on several factors. The reliable recognition rates, the reduction of the calculation time and the robustness are decisive assets. In the approach we propose, we aim to satisfy these factors. This work is divided into three phases:

In the first phase, we will address two major points that will later be used to build our final system. The first point during this phase is to show that the recognition system based on local feature extraction with a compound local binary pattern (CLBP) can provide reliable results.

The second point concerns the matching phase, which is a very important step and can be costly in computation time. To satisfy this constraint, we will proceed with distance-based classifiers to reduce the computation time. We are going to test three classifiers during the first phase and take the most appropriate one in relation to the local extraction method used (recognition rate and computation time). The system used for this evaluation is shown in Fig. 1.

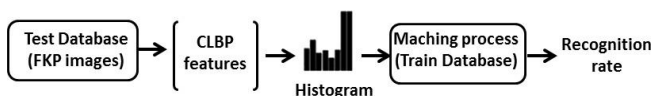


Fig. 1. FKP Recognition System Adopted for Evaluation.

During the second phase, we will proceed to an analysis and a comparison of the results obtained with other methods, in order to demonstrate the efficiency of the method to be used in our system.

In the end, the last phase consists in using the results and the conclusions obtained in the two previous phases to test the effectiveness of our global system, which is based on the concept of fusion at score level. The results obtained in the last phase will be analyzed and compared with phase 2. The proposed FKP recognition system is shown in Fig. 2.

### A. Local Feature Extraction Process

Recognition systems generally satisfy two overriding conditions: high recognition rates and low computational cost. Note that the capture phase may be affected by environmental conditions. To overcome this problem, the local feature extraction phase will be ensured by a robust variant of Local Binary Patterns (LBP), this variant is called Compound Local Binary Patterns (CLBP) [25]. Ojala and Ai [27] introduced the local binary pattern method for the first time as an efficient method for feature extraction from images. The features extracted by this process have provided an efficient means for texture segmentation and classification.

The local binary pattern is recognized by a gray scale texture operator characterizing the local spatial structure of the image texture [28]. Given a central pixel in the image, a pattern code is calculated by comparing it to its neighbors. This procedure is illustrated in Fig. 3.

The LBP operator takes the form:

$$LBP(x_c, y_c) = \sum_{n=0}^7 2^n S(i_n - i_c) \tag{1}$$

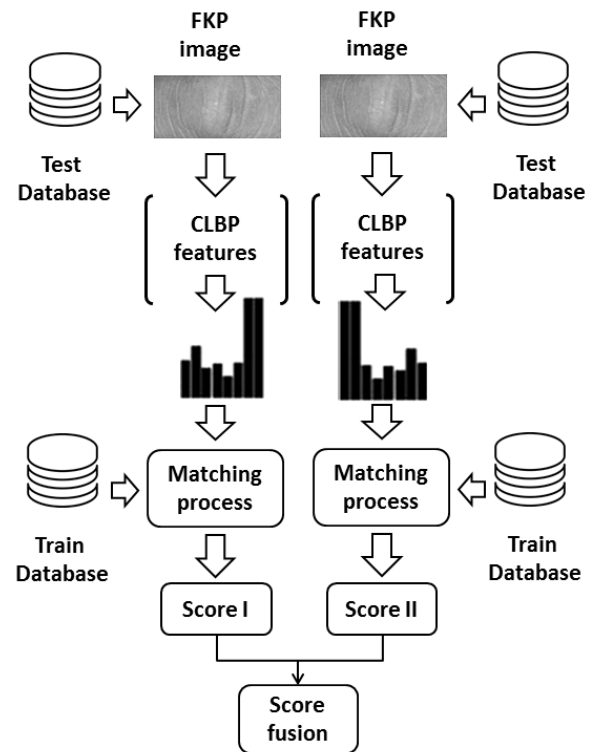


Fig. 2. The Proposed FKP Recognition System.

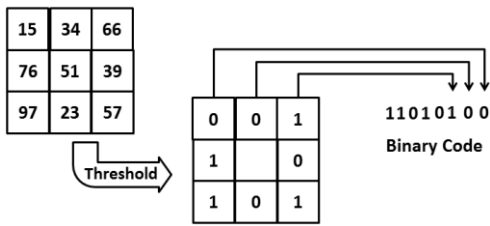


Fig. 3. Local Binary Pattern Operator.

where in this case  $n$  runs over the eight neighbors of the central pixel  $c$ ,  $i_c$  and  $i_n$  are the gray-level values at  $c$  and  $n$ . Function  $S(x)$  is shown below,

$$S(x) = \begin{cases} 1 & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases} \quad (2)$$

The LBP operator employs a process that relies solely on the use of the sign of the difference between two gray values, which often leads to a failure to generate binary codes consistent with the texture properties for local region. To overcome this problem, we will use an extension of this operator. This extension assigns a 2P bit code to the central pixel based on the gray values of the local neighborhood comprising P neighbors; this extension is the Compound Local Binary Pattern method (CLBP) [29]. The CLBP operator employs two bits for each neighbor in order to encode the sign as well as magnitude information of the difference between the center and the neighbor gray values, unlike the LBP that uses only one bit for each neighbor by representing the sign of the difference between the center and the corresponding neighbor gray values. In this case, the first bit is representing the sign of the difference between the center and the corresponding neighbor gray values as the basic LBP encoding. The second bit is for encoding the magnitude of the difference with respect to a threshold value, which is the average magnitude  $M_{avg}$  of the difference between the center and the neighbor gray values in the local neighborhood of interest. This CLBP operator chooses the value of 1 for the second bit if the magnitude of the difference between the center and the corresponding neighbor is greater than the threshold  $M_{avg}$ . Other way, it takes the value of 0. Thus, the indicator  $s(x)$  of equation 2 is replaced by the following function:

$$s(i_p, i_c) = \begin{cases} 00 & i_p - i_c < 0, |i_p - i_c| \leq M_{avg} \\ 01 & i_p - i_c < 0, |i_p - i_c| > M_{avg} \\ 10 & i_p - i_c \geq 0, |i_p - i_c| \leq M_{avg} \\ 11 & \text{otherwise} \end{cases} \quad (3)$$

where  $i_c$  and  $i_p$  are the gray values of the central pixel and the neighbors, and the average magnitude of the difference between  $i_p$  and  $i_c$  in the local neighborhood is  $M_{avg}$ . The illustration of the CLBP operator is shown in the Fig. 4.

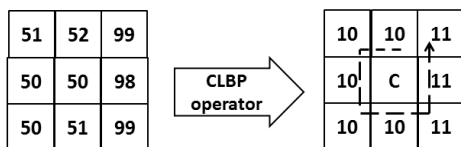


Fig. 4. Compound Local Binary Pattern Operator.

In Fig. 4, it can be observed that the CLBP operator discriminates the neighbors of the northeast, east and southeast directions because they have higher gray values than the other neighbors, thus producing a consistent local model.

### B. Matching Process

Our choice of classifiers that can be used in our system was based on two important points. The first concerns the computation time, which is an important asset for the success of a recognition system, which directs us towards distance-based classifiers. The second point concerns the resolution of our images, which is not high, so we avoid classifiers such as SVM, which is rather oriented for high resolutions [30], [31]. Taking into account the guidelines already mentioned, we opted for a set that includes three distance-based classifiers, which we will experiment with the extraction method employed and see their performance for recognition rates generation. These classifiers are based on the Euclidean distance, Jeffrey Divergence and City-block.

The Euclidean distance is the most common distance metric used for low dimensional data sets, examines the root of square differences between the coordinates of a pair of objects. This is most generally known as the Pythagorean Theorem. For testing we used this classifier, for calculating the minimum distance between the test image and train image. Euclidean distance  $d$  is presented as follows:

$$d(x, y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2} \quad (4)$$

The Jeffrey divergence is a modification of the Kullback-Leibler (KL) divergence, if  $P = (p_1, \dots, p_N)$  and  $Q = (q_1, \dots, q_N)$  are two discrete distributions, the Jeffrey divergence between  $P$  and  $Q$  is defined as:

$$D(P, Q) = \sum_i (p_i \log \frac{p_i}{m_i} + q_i \log \frac{q_i}{m_i}) \quad (5)$$

$$\text{Where } m_i = \frac{(p_i + q_i)}{2}$$

The city-block distance classifier, Manhattan distance classifier, also called, rectilinear distance, L1 distance, L1 norm, Manhattan length. It represents the distance between points in a city road grid. It examines the absolute differences between the coordinates of a pair of objects as follows:

$$d(x, y) = \sum_{i=1}^n |x_i - y_i| \quad (6)$$

## III. EXPERIMENTAL RESULT

In this section, we will proceed with various experiments to prove the reliability of the proposed recognition scheme. For a comparative evaluation of our recognition system, we will conduct experiments on the PolyU database [26]. This database is part of the databases, which can be described as referential databases in this field.

### A. FKP PolyU Database

The PolyU FKP database is identified as a reference in biometrics research work (Fig. 5). This benchmark is used for evaluating the performance of the majority of FKP recognition systems that have been studied. The database construction was made thanks to the participation of 165 volunteers, including 40 women and 125 men. Among them, 143 individuals aged

between 20 and 30 years old and the rest between 30 and 50 years old. Two separate sessions have been designed to collect FKP images. During these two sessions, the volunteer is asked to provide six images for each of the right index finger, left index finger, right middle finger and left middle finger. All the original FKP images used have a resolution equal to 220x110. In the end, each subject is determined by  $12 \times 4 = 48$  images, 12 FKP images of each finger. Thus, the database total is 7920 images from 660 fingers.

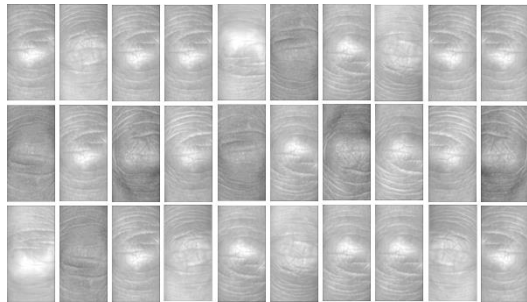


Fig. 5. Finger Knuckle Print PolyU Database.

### B. Evaluation and Analysis of Results

In order to validate our recognition system, we have established an evaluation process divided into three different phases. The first step concerns a single finger knuckle print evaluation of our scheme, this step aims to determine the classifiers able to provide reliable results, and then we choose the most efficient ones. During this step, we will perform detailed experiments for each finger. The object of these experiments will be to determine the most suitable classifiers for our case study (recognition rate), the overall time taken by the matching phase for all the classes which are of the order of 165 classes, each class uses 6 test images which will be compared to  $6 \times 165$  images = 990 images, we will call later in the experiments this computation time: Matching Process Time MPT. We will also introduce the division of the image into sub-images according to the type of resolution (square or rectangular) and conclude whether it is possible to introduce this kind of rectangular division, which is generally absent in the literature. In the second step, after having obtained the results and their analysis, we will proceed to a comparison with the other mechanisms already cited in the literature, which have used finger knuckle prints for the creation of recognition systems and conclude on the obtained performance. Finally, in the last part, we will use the directives obtained in our global approach, see its impact on performance and demonstrate the reliability of the proposed approach.

Single finger knuckle print evaluation in this first step, we will apply the same experimental protocol used by the others systems cited in literature. The 6 images captured during the first session are used to create the training database and the 6 images captured in the second session for the testing database. Therefore, for each volunteer, there are six training samples and six testing samples (Fig. 6). Our approach is based on local methods. The performance of scheme is evaluated with different sizes of sub images for each FKP image.

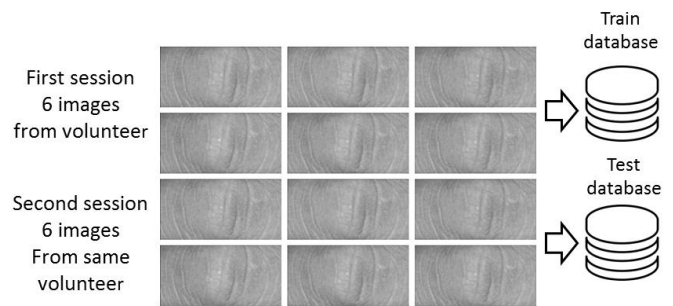


Fig. 6. Standard Protocol used for FKP Experiment.

We have categorized block sizes into three divisions: large division, medium division and small division. Large division is defined by two type of size block: 110x110 and 64x64, for medium size: 48x48 and 32x32 pixels, and for small size: 24x24 and 16x16. In order to increase Classification Process Performance of our approach, we have inspected many classifiers based on the squared Euclidean distance, the divergence of Jeffrey and City-Block. The recognition rates for each finger with the different divisions of sub-images are presented in the comparative Tables I, II, III, IV, V, VI, VII and VIII. This comparative evaluation is made with the intention of showing the most adaptive classifiers in our case.

1) *Result of experiment on left index finger:* The Table I shows the adequacy of classifiers based on distance city-Block and Jeffery divergence compared to the classifier based on Euclidean distance. The city-block classifier gives the best recognition rate value 98.18% for divisions (sub-images) of 16x16 pixels. We also notice that the Matching Process Time MPT increases proportionally with the decrease of the dividing block size. This observation is normal, since the smaller the block size, the more the number of sub-images increases (for image), and therefore the histogram of the image too. As we can also notice that even if the recognition rates can be equivalent between city-block and Jeffrey divergence, the MPT with the Jeffrey divergence classifier is much higher, we can cite the case of the 16x16 pixel block where the MPT for city block equal to 21.30s while that obtained with Jeffrey divergence equal to 852.96s.

It should be noted that these divisions (16x16, 24x24, 32x32, 48x48 and 64x64) are the most used in the literature. Nevertheless, we must not limit ourselves to this rule often used by researchers; we can extend the division with blocks of rectangular sub-images and not only square ones. In our case, we have an image whose size is 220x110 pixels; it would be wise to choose a division, which participates in an optimal construction of the histograms of each sub-image. In this sense to verify this proposal and based on the analysis of the optimal results obtained with blocks, which vary between 16x16 and 24x24 pixels and the shapes of the lines in the FKP images, we will extend the experiment to the 11x22 pixels blocks. This block division has exact multiples for the size of our FKP (220x110) images  $11 \times 20 = 220$  pixels and  $22 \times 5 = 110$  pixels, which gives for the division  $(220 \times 110) \setminus (11 \times 22) = 100$  sub-images.

TABLE I. RECOGNITION RATE FOR LEFT INDEX WITH TYPICAL BLOCKS

		Table Recognition rate RR (Left index)		
Block size	RR/MPT	Euclidian distance	Jeffrey divergence	City-Block
110x110	RR	78,69%	<b>88,79%</b>	87,07%
	MPT	1,383645 s	22,262518	1,678907 s
64x64	RR	90,10%	<b>94,75%</b>	93,74%
	MPT	1,525478 s	73,938122 s	2,721189 s
48x48	RR	92,12%	95,35%	<b>95,55%</b>
	MPT	1,142311 s	96,140630 s	2,851411 s
32x32	RR	94,95%	<b>96,86%</b>	96,67%
	MPT	1,932484 s	273,279155 s	7,005333 s
24x24	RR	96,16%	97,37%	<b>98,08%</b>
	MPT	2,615035 s	422,307223 s	10,967445 s
16x16	RR	96,26%	97,98%	<b>98,18%</b>
	MPT	3,784349 s	852,959212 s	21,302479 s

TABLE II. RECOGNITION RATE FOR LEFT INDEX WITH RECTANGULAR BLOCK

		Table Recognition rate RR (Left index)		
Block size	RR/MPT	Euclidian distance	Jeffrey divergence	City-Block
11x22	RR	96,56%	97,98%	<b>98,48%</b>
	MPT	3,613546 s	877,053453 s	21,665596 s

In the Table II, we notice that the division of the image into sub-images with blocks size 11x22 pixels gives an optimal recognition rate with a value of 98.48%, this value rivals that obtained by the 16x16 block whose value is 98,18%. We are going to opt for this additional experience for the rest of the fingers (right Index, left middle and right middle).

2) *Result of experiment on right index finger:* In Table III, the obtained results affirm once again the adequacy of the resulting recognition rate with the classifiers based on the Jeffrey divergence and the City-block. We note that the value of the high recognition rate obtained is 98.48% with city block for a size block 24x24 pixels. As we mentioned before we are going to continue the complement of the experiment with the block whose size is 11x22 pixels. For MPT, we observe the same remark as before.

In the Table IV, we still notice the same remark and that the division of the image into sub-images with block of size 11x22 pixels gives optimal recognition rate with a value of 98.69%, this value remains equivalent or higher than that obtained by the 24x24 block whose value is 98.48%. It should be noted that even if we have equivalence in terms of recognition rates for the 11x22 pixels blocks with Jeffrey divergence and City-block, there remains the MPT factor, which gives a considerable advantage for City-block. The operation with Jeffrey divergence is more cost in computation time, MPT equal: 570, 16 s.

3) *Result of experiment on left middle finger:* Table V still shows the results superiority of the obtained recognition rates with the classifiers based on the Jeffrey and City-block divergence. We note that the value of the highest recognition rate is 99.29% with city-block for size 16x16 pixels. We will continue our experiments with the complement concerning the 11x22 pixels blocks.

TABLE III. RECOGNITION RATE FOR RIGHT INDEX WITH TYPICAL BLOCKS

		Table Recognition rate RR (Right index)		
Block size	RR/MPT	Euclidian distance	Jeffrey divergence	City-Block
110x110	RR	79,90%	<b>90,80%</b>	88,89%
	MPT	0,938604 s	14,560413 s	1,169245 s
64x64	RR	91,61%	<b>96,57%</b>	95,86%
	MPT	1,049516 s	47,949257 s	1,799466 s
48x48	RR	93,94%	<b>97,37%</b>	97,27%
	MPT	1,153916 s	94,779885 s	2,828075 s
32x32	RR	96,77%	97,88%	<b>98,28%</b>
	MPT	1,378431 s	175,976525 s	175,976525 s
24x24	RR	96,36%	98,18%	<b>98,48%</b>
	MPT	1,841294 s	286,073108 s	7,029212 s
16x16	RR	96,36%	<b>97,88%</b>	<b>97,88%</b>
	MPT	2,416625 s	554,749914 s	13,857284 s

TABLE IV. RECOGNITION RATE FOR RIGHT INDEX WITH RECTANGULAR BLOCK

		Table Recognition rate RR (Right index)		
Block size	RR/MPT	Euclidian distance	Jeffrey divergence	City-Block
11x22	RR	96,46%	98,69%	98,69%
	MPT	2,451808 s	570,162648 s	14,020781 s

TABLE V. RECOGNITION RATE FOR LEFT MIDDLE WITH TYPICAL BLOCKS

		Table Recognition rate RR (Left middle)		
Block size	RR/MPT	Euclidian distance	Jeffrey divergence	City-Block
110x110	RR	82,12%	<b>91,41%</b>	90,61%
	MPT	0,942734 s	14,488085 s	1,122162 s
64x64	RR	91,81%	<b>95,86%</b>	<b>95,86%</b>
	MPT	1,035171 s	47,818859 s	1,788124 s
48x48	RR	94,24%	<b>97,07%</b>	<b>97,07%</b>
	MPT	1,163299 s	95,044139 s	2,828460 s
32x32	RR	96,76%	97,68%	<b>98,48%</b>
	MPT	1,359258 s	175,401198 s	4,502957 s
24x24	RR	97,58%	98,59%	<b>98,89%</b>
	MPT	1,726323 s	287,939161 s	7,067663 s
16x16	RR	97,17%	98,89%	<b>99,29%</b>
	MPT	2,704647 s	550,270049 s	13,806499 s

TABLE VI. RECOGNITION RATE FOR LEFT MIDDLE WITH RECTANGULAR BLOCK

		Table Recognition rate RR (Left middle)		
Block size	RR/MPT	Euclidian distance	Jeffrey divergence	City-Block
11x22	RR	96,87%	<b>98,99%</b>	98,89%
	MPT	2,508779 s	569,049885 s	14,005159 s

The results obtained in the Table VI show that the division of the image into sub-images with 11x22 pixels always gives optimal recognition rates with a value of 98.89% in the case of city-block. This value still remains near to that obtained by the 16x16 pixels division, which has a value equal to 99.29% and equal to that obtained by the 24x24 pixels division.

4) *Result of experiment on right middle finger:* Table VII keeps the same conclusion made in the three previous experiments, the superiority of the recognition rates obtained with the classifiers based on the Jeffrey divergence and City-block is maintained. We see that the value of the most optimal recognition rate is 98.89% with city-block for the size 24x24 pixels. We are going to finish the experiments of this first part with the complement concerning the 11x22 block as before.

The block sizes of 11x22 pixels still ensure optimal recognition rates in Table VIII with a value of 98.89% for the 11x22 pixel block. This value is equal to the highest obtained in Table VII by the 24x24 pixel block with a value of 98.89%. In the end, we can conclude that dividing the sub-images into rectangular blocks can give results as optimal as square blocks.

TABLE VII. RECOGNITION RATE FOR RIGHT MIDDLE WITH TYPICAL BLOCKS

		Table Recognition rate RR (right middle)		
Block size	RR/MPT	Euclidian distance	Jeffrey divergence	City-Block
110x110	RR	83,23%	<b>92,63%</b>	90,71%
	MPT	0,950980 s	14,532302 s	1,141061 s
64x64	RR	93,03%	<b>96,67%</b>	<b>96,67%</b>
	MPT	1,019915 s	47,960191 s	1,817366 s
48x48	RR	94,54%	96,77%	<b>97,37%</b>
	MPT	1,665615 s	146,035030 s	4,346636 s
32x32	RR	97,27%	97,78%	<b>98,48%</b>
	MPT	1,935565 s	271,733796 s	6,986675 s
24x24	RR	97,07%	98,59%	<b>98,89%</b>
	MPT	2,396355 s	438,833175 s	10,932720 s
16x16	RR	96,77%	<b>98,38%</b>	98,18%
	MPT	3,809532 s	843,720436 s	21,210157 s

TABLE VIII. RECOGNITION RATE FOR RIGHT MIDDLE WITH RECTANGULAR BLOCK

		Table Recognition rate RR (right middle)		
Block size	RR/MPT	Euclidian distance	Jeffrey divergence	City-Block
11x22	RR	97,37%	98,48%	<b>98,89%</b>
	MPT	3,584845 s	868,403304 s	21,696460 s

### C. Comparison and Analysis of Results

As we have already mentioned, we have set ourselves as objectives, the construction of a robust recognition system which offers a high recognition rate and a reduced computation time. The results obtained confirm our choice of distance-based classifiers with the method used for the extraction of local characteristics. Our choice thereafter to ensure the comparison of our mechanism with others in the literature will be based on the use of the city-block distance. The optimal results are in the following table:

TABLE IX. OPTIMAL OBTAINED RECOGNITION RATE

Table optimal Recognition rate				
Block size	Left index	Right index	Left middle	Right middle
11x22	<b>98,48%</b>	<b>98,69%</b>	98,89%	<b>98,89%</b>
16x16	98,18%	97,88%	<b>99,29%</b>	98,18%
24x24	98,08%	98,48%	98,89%	<b>98,89%</b>

We notice on the Table IX, that the block which shows the highest rates on average is the 11x22 block; we will take its results and compare them with those of the literature to remove the performance of the adopted mechanism.

TABLE X. COMPARATIVE STUDY

Table optimal Recognition rate				
Methods	Left index	Right index	Left middle	Right middle
PCA+LDA [22]	50.64 %	47.00%	51.08 %	54.68 %
CLPP [24]	86.58 %	86.43 %	85.89 %	86.16 %
OCLPP [24]	87.87 %	87.49 %	86.94 %	87.38 %
MSLBP [23]	93.80 %	94.70 %	92.20 %	94.80 %
LGBP [33]	94.14%	94.24%	97.27%	94.75%
LBP+DCT [32]	98.2%	98%	98.7%	97.1%
<b>Our work</b>	<b>98,48%</b>	<b>98,69%</b>	<b>98,89%</b>	<b>98,89%</b>

Table X shows that the preliminary results of the adopted process claim to be reliable, however there are still other aspects that we want to address to study and improve the support of our approach. This is what we will see in the next phase.

### D. Global Evaluation of Proposed Approach

In this section, we will conduct our experiments to evaluate the proposed approach. These experiments consist in exploiting the sizes of the blocks, which have shown their performance previously, with a single finger knuckle print (24x24, 16x16, and 11x22). In this evaluation, we will use the approach with a fusion at the score level, with the city-block distance for the set of multi-instance combinations: left index with left middle (LI, LM) and right index with right middle (RI, RM). This choice is due to the fact that these combinations belong to the same hand, which facilitates the use and creation of sensors in a real recognition system. We will use the Cumulative Matching Characteristics (CMC) curves for each fusion case to measure the identification accuracy. CMC curves demonstrate the ability of a recognition system to identify a given user in a set of data. As the CMC



curve decreases, this represents an increasing amount of impostor images that are more similar than images of the required class, otherwise performance increases.

1) *Results of fusion left index and left middle results for block size 24x24 pixels:* In Fig. 7, we notice that the resulting curve for the fusion of the LI+LM instances is much higher than the other curves and it quickly tends towards 1. The higher recognition rate obtained with the approach equal to 99.90%, this rate is higher than the best rate obtained for studied systems with single instance, in this case left and right middle with 98.89%.

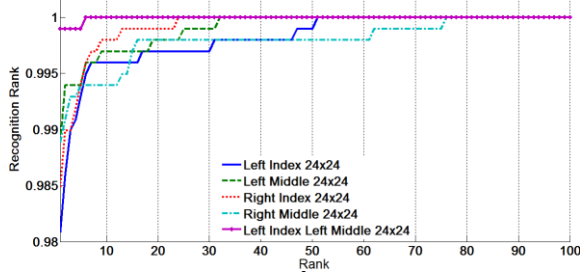


Fig. 7. CMC Curve for Fusion LI+LM with 24x24 Blocks.

a) *Results for block size 16x16 pixels:* In Fig. 8, we will report the same remark in the case of the block equal to 16x16. The recognition rate obtained with the approach is equal to 99.90%; this rate is higher than the best rate obtained for the systems studied previously.

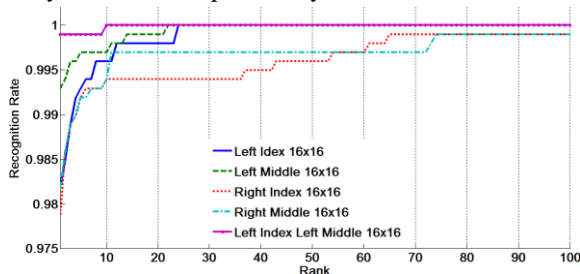


Fig. 8. CMC Curve for Fusion LI+LM with 16x16 Blocks.

b) *Results for block size 11x22 pixels:* In Fig. 9, we used the 11x22 pixel blocks and the resultant is a perfect curve with a score of 100% which outperforms all the curves.

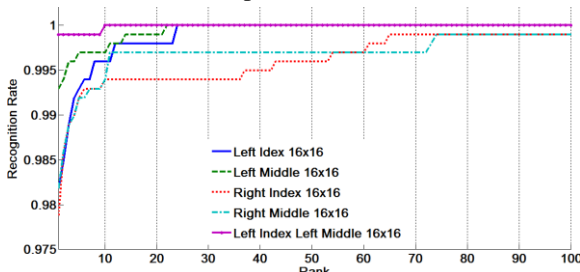


Fig. 9. CMC Curve for Fusion LI+LM with 11x22 Blocks.

2) *Results of fusion right index and right middle*

a) *Results for block size 24x24 pixels:* In Fig. 10, despite the recognition rate that the approach offers and which is equal = 99.80% for the fusion between Right Index and

Right Middle instances (RI, RM), but we notice the Right Index (system with single instance) curve overlaps with the (RI, RM) curve and it reaches values 1 well before the melting curve and this is due to the inter-class similarities.

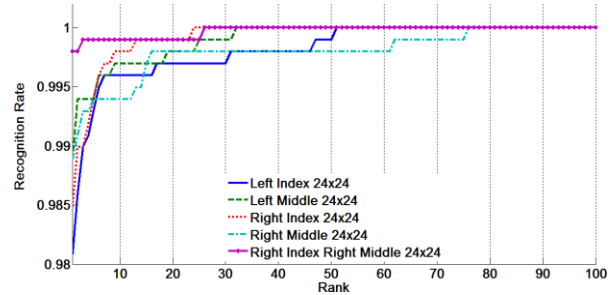


Fig. 10. CMC Curve for Fusion RI+RM with 24x24 Blocks.

b) *Results for block size 16x16 pixels:* In Fig. 11, despite the recognition rate offered by the approach and which is equal = 99.70% for the fusion between the instances Right Index and Right Middle (RI, RM) with the 16x16 blocks, but we notice that the curve (RI, RM) on the first 100 ranks it does not reach the values 1 and that it is exceeded by the Left Index curve and the Left Middle curve towards rank 24.

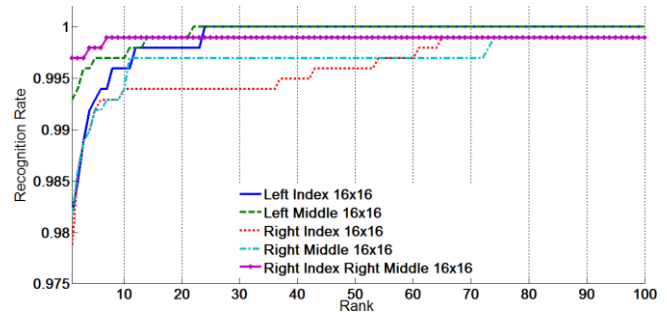


Fig. 11. CMC Curve for Fusion RI+RM with 16x16 Blocks.

c) *Results for block size 11x22 pixels:* In Fig. 12, we notice that the resulting curve for the fusion of the right index and right middle instances, in the case where the block size equal to 11x22 pixels is much higher than the other curves. The recognition rate obtained with the approach is equal to 99.70%, and reaches quickly the value 1 quickly before others curves.

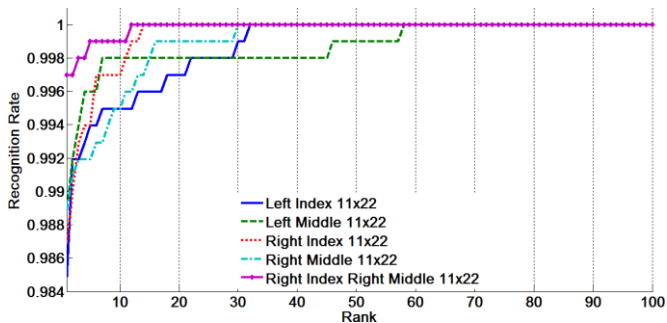


Fig. 12. CMC Curve for Fusion RI+RM with 11x22 Blocks.

3) *Comparison between fusion curves (LI, LM) and (RI, RM):* The results obtained show that the choices do not depend on the size of the block but also on its most suitable shape for



the subdivision of the source images. The performance of the results is shown in the Table XI. To confirm these hypotheses, we will obtain the CMC curves of all the mergers (LI, LM) and (RI, RM) with the 16x16, 24x24 and 11x22 blocks.

TABLE XI. RESULTS OF OUR APPROACH

	Recognition rate		
	Block size		
Instances	24x24	16x16	11x22
LI+LM	99,90%	99,90%	100%
RI+RM	99,80%	99,70%	99,70%

a) Results for fusion(LI, LM): In Fig. 13, The curve that represents the LI+LM fusion with the 11x22 block is a curve that surpasses the 2 other fusion curves with the 16x16 and 24x24 blocks. We notice that the fusion curve with the 24x24 blocks is more efficient and its convergence is faster towards the 1 than the 16x16 curve.

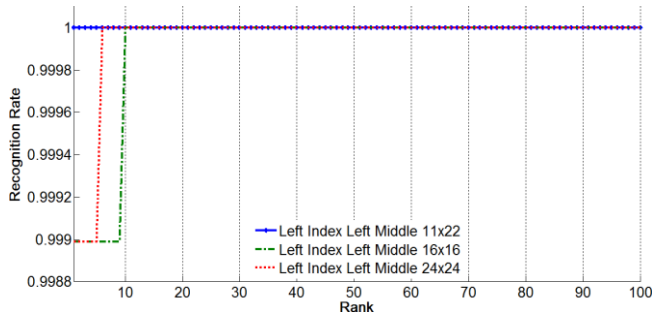


Fig. 13. CMC Curves for Fusion LI+LM.

b) Results for fusion(RI, RM): In Fig. 14, although the RI+RM fusion curve with the 24x24 block begins with a higher recognition rate compared to the RI+RM fusion curve with the 11x22 block, the latter tends towards ones more quickly than the curve 24x24 and 16x16.

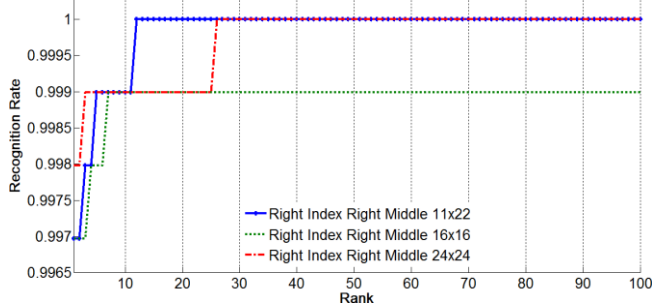


Fig. 14. CMC Curves for Fusion RI+RM.

All the experiments show very satisfactory results with the fusion approach adopted. The curves in Fig. 7, 8, 9, 10, 11 and 12 support these results. The introduction of the notion of the "block size \ image resolution" ratio in the experiments has shown its effectiveness and its ability to improve the results already obtained. The CMC curves Fig. 13 and Fig. 14 clearly demonstrate this improvement.

#### IV. CONCLUSIONS

In this paper, we evaluated the performance of the local CLBP descriptor, the influence of the block size parameter and its shape on the recognition rate. To improve efficiency and accuracy, we proposed an approach based on multi-instance fusion at the score level. The experimental results on the PolyU FKP reference database clearly show that the proposed approach increases the recognition rates (between 99.70% and 100%) and that it reduces the influence on the variance of the rates by taking charge of the adequate divider block according to the resolution of the image for the optimal construction of the histograms. Thus, we can conclude that this approach provides a noticeable performance improvement and can be usefully used for FKP recognition systems. The future works will focus on improving the security side of the recognition systems construction based on hand modalities. This improvement will aim to reduce the possibility of personal identity theft, while reducing the complexity of the mechanism to be built.

#### REFERENCES

- [1] G. E. Mardini, S. Massari, Body, biometrics and identity, Bioethics 22 (2008) p. 488–498.
- [2] Mahalakshmi B S and Sheela S V, "A Novel Feature Extraction for Complementing Authentication in Hand-based Biometric" International Journal of Advanced Computer Science and Applications(IJACSA), 12(9), 2021.
- [3] E. Marasco and A. Ross, "A survey on antispoofing schemes for fingerprint recognition systems," ACM Computing Surveys (CSUR), vol. 47, p. 28, 2015.
- [4] Zhao, T.; Liu, Y.; Huo, G.; Zhu, X. A deep learning iris recognition method based on capsule network architecture. IEEE Access 2019, 7, 49691–49701.
- [5] Rachida Tobji, Wu Di, Naem Ayoub and Samia Houassi, "Efficient Iris Pattern Recognition Method by using Adaptive Hamming Distance and 1D Log-Gabor Filter" International Journal of Advanced Computer Science and Applications(IJACSA), 9(11), 2018.
- [6] Zhang L, Shen Y, Li H Y, Lu J W, "3D palmprint identification using block-wise features and collaborative representation", IEEE Transactions on Pattern Analysis and Machine Intelligence, 37(8) :1730-1736, 2015.
- [7] Zhang D, Automated Biometrics: Technologies and Systems, Dordrecht: Kluwer Academic, 2000.
- [8] E. Yrk, E. Konukoglu, B. Sankur , and J. Darbon, "Shape-Based Hand Recognition," IEEE Transaction On. Image Processing, Vol. 15, No. 7, page 1803-1815, 2006.
- [9] D. N. Fernández, "Development of a hand pose recognition system on an embedded computer using Artificial Intelligence," 2019 IEEE XXVI International Conference on Electronics, Electrical Engineering and Computing (INTERCON), 2019, pp. 1-4.
- [10] .D. Zhang, W. Zuo, F. Yue, A comparative study of palmprint recognition algorithms, ACM Computing Surveys 44 (1) (2012) 2:1–37.
- [11] Jia W, Zhang B, Lu J T, Zhu Y H, Zhao Y, Zuo W M, Ling H B, Palmprint recognition based on complete direction representation. IEEE Transactions on Image Processing, 26(9) : 4483-4498, 2017.
- [12] Alonso-Fernandez, F., Fierrez, J., Ortega-Garcia, J., Gonzalez-Rodriguez, J., Fronthaler, H., Kollreider, K., Bigun, J.: A comparative study of fingerprint image quality estimation methods. IEEE Trans. on Information Forensics and Security 2(4), 734–743 (2007).
- [13] Nogueira, R.F., de Alencar Lotufo, R., and Machado, R.C.: 'Fingerprint liveness detection using convolutional neural networks', IEEE transactions on information forensics and security, 2016, 11, (6), pp. 1206-1213.

- [14] Z. Feng, B. Yang, Y. Chen, Y. Zheng, T. Xu, Y. Li, T. Xu, D. Zhu, Features extraction from hand images based on new detection operators, *Pattern Recognition* 44 (5) (2011) 1089–1105.
- [15] N. Duta, A survey of biometric technology based on hand shape, *Pattern Recognition* 42 (11) (2009) 2797–2806.
- [16] Tertychnyi, P., Ozcinar, C., and Anbarjafari, G.: ‘Low-quality fingerprint classification using deep neural network’, *IET Biometrics*, 2018, 7, (6), pp. 550–556.
- [17] Matsumoto, T., Matsumoto, H., Yamada, K. and Hoshino, S., Impact of artificial gummy fingers on finger-print systems, *Proc. SPIE, Optical Security and Counterfeit Deterrence Techniques IV*, vol. 4677, pp. 275–289, January 2002.
- [18] A. Kumar, C. Ravikanth, Personal authentication using finger knuckle surface, *IEEE Transactionson Information Forensics and Security* 4 (1) (2009) 98–109.
- [19] L. Zhang, L. Zhang, D. Zhang, and Z. Guo, “Phase congruency induced local features for finger-knuckle-print recognition,” *Pattern Recognition*, vol. 45, no. 7, pp. 2522–2531, 2012.
- [20] A. Kumar, C. Ravikanth, Personal authentication using finger knuckle surface, *IEEE Transactionson Information Forensics and Security* 4 (1) (2009) 98–109.
- [21] L. Zhang, L. Zhang, D. Zhang, and H.L. Zhu. Online finger-knuckle print verification for personal authentication. *Pattern Recognition*, 43(7):2560–2571, July 2010.
- [22] Z. S. Shariatmadar and K. Faez, “An efficient method for fingerknuckle-print recognition by using the information fusion at different levels,” in 2011 International Conference on Hand-Based Biometrics (ICHB), 2011, pp. 1–6.
- [23] W. El-Tarhouni, M. Shaikh, L. Boubchir, and A. Bouridane, “Multiscale shift local binary pattern based-descriptor for finger-knuckleprint recognition,” in *Microelectronics (ICM)*, 2014 26th International Conference on, pp. 184–187, Dec 2014.
- [24] X. Jing, W. Li, C. Lan, Y. Yao, X. Cheng, and L. Han, “Orthogonal complex locality preserving projections based on image space metric for finger-knuckle-print recognition,” in 2011 International Conference on Hand-Based Biometrics (ICHB), 2011, pp. 1–6.
- [25] Faisal Ahmed, Emam Hossain, A.S.M. Hossain Bari and ASM Shihavuddin, “Compound Local Binary Pattern (CLBP) for Robust Facial Expression Recognition,” *IEEE International Symposium on Computational Intelligence and Informatics*, pp.391-395, Budapest, Hungary, 2011.
- [26] PolyU finger knuckle print database (FKP) <http://www.comp.polyu.edu.hk/biometrics/FKP.htm>.
- [27] T. Ojala, M. Pietikainen and D. Harwood, A comparative study of texture measures with classification based on feature distributions, *Pattern Recognition*, vol. 29, 1996.
- [28] ] Stan Z. Li, ChunShui Zhao, XiangXin Zhu, Zhen Lei, 2D+3D Face Recognition by Fusion at Both Feature and Decision Levels, In *Proceedings of IEEE International Workshop on Analysis and Modeling of Faces and Ges-tures*. Beijing. Oct 16, 2005.
- [29] AHMED, Faisal, HOSSAIN, Emam, BARI, A. S. M. H., et al. Compound local binary pattern (clbp) for rotation invariant texture classification. *International Journal of Computer Applications*, 2011, vol. 33, no 6, p. 5-10.
- [30] J. Inglada, Automatic recognition of man-made objects in high resolution optical remote sensing images by SVM classification of geometric image features, *ISPRS Journal of Photogrammetry and Remote Sensing*, vol.62, issue.3, pp.236-248, 2007.
- [31] ] Giannis Lantzanakis, Zina Mitra, Nektarios Chrysoulakis, X-SVM: An Extension of C-SVM Algorithm for Classification of High-Resolution Satellite Imagery. *IEEE Trans. Geosci. Remote. Sens.* 59(5): 3805–3815 (2021).
- [32] M. Amraoui, M. El Aroussi, R. Saadane, M. Wahbi. Finger-knuckle-print recognition based on local and global feature sets. In *Journal of Theoretical and Applied Information Technology*, pp 054–060 Vol. 46. No. 1 – 2012.
- [33] Xiong, M., Yang,W., Sun, C.: Finger-knuckle-print recognition using lgbp. In: *International Symposium on Neural Networks*. ISSN (2011) 270–277.

# Detection of Credit Card Fraud using a Hybrid Ensemble Model

Sayali Saraf<sup>1</sup>, Anupama Phakatkar<sup>2</sup>  
Department of Computer Engineering  
SCTR's Pune Institute of Computer Technology  
Pune, India

**Abstract**—The rising number of credit card frauds presents a significant challenge for the banking industry. Many businesses and financial institutions suffer huge losses because card users are reluctant to use their cards. A primary goal of fraud detection is to identify prior transaction patterns to detect future fraud. In this paper, a hybrid ensemble model is proposed to combine bagging and boosting techniques to distinguish between fraudulent and legitimate transactions. During the experimentation two datasets are used; the European credit card dataset and the credit card stimulation dataset which are highly imbalanced. The oversampling method is used to balance both datasets. To overcome the problem of unbalanced data oversampling method is used. The model is trained to predict output results by combining random forest with Adaboost. The proposed model provides 98.27 % area under curve score on the European credit cards dataset and the stimulation credit card dataset gives 99.3 % area under curve score.

**Keywords**—Credit card; hybrid ensemble model; bagging; boosting; data imbalance

## I. INTRODUCTION

There is a growing issue of financial fraud in the government, businesses, and financial sector with significant implications [1]. In credit card fraud, purchases occur on a cardholder's account without the cardholder's knowledge or consent. It is crucial to prevent fraud by taking all necessary precautions when carrying out these transactions. Bank regulators must also employ snipping technology to anticipate these thefts. Predicting the transactions that account holders will make but which will be completed by other people with access to the account is a fraud detection method for our dataset. It is a complex issue that needs to be resolved by both the account holder and the bank so that other customers don't face the same issue. However, there is a problem of class inequality with this issue. An individual consumer will complete many more legitimate transactions than fraudulent ones, or even none at all. A transaction that differs from a customer's previous purchases might be considered fraud. As credit card transactions increase in popularity for payment, academics are focusing on several strategies for fighting credit card fraud. The most common yet challenging issue is credit card fraud detection. As a result of the limited amount of credit card data, it is challenging to match a pattern for a dataset. Second, many records in the collection could include fraudulent transactions that follow a pattern of honest activity [2]. There are also some limitations to the issue. Firstly, study results are often classified and regulated, making them

unavailable. Additionally, classified data sets are not readily available to the general public. Due to this, benchmarking specific models may be challenging. It is also difficult to develop solutions due to the security issue, which limits the exchange of concepts and techniques for detecting fraud, particularly credit card fraud [3]. The last point is that data sets are continually changing and evolving. It produces profiles of legitimate and fraudulent behavior separate from current valid transactions that may have been fraudulent in the past. In this paper, we will use a variety of machine learning algorithms, including logistic regression, random forest, and AdaBoost, to evaluate the performance of our proposed model. Two credit card datasets are used in the experiment, one of which is very skewed and unbalanced. The hybrid ensemble model is used to differentiate between fraudulent and legal transactions.

The work presented in the paper can be summarized as follows:

- 1) A hybrid ensemble model is proposed to classify fraudulent and legitimate transactions. The system uses an Adaboost, random forest, and Logistic regression to build a classifier.
- 2) The oversampling method and the removing outliers' approach are two methods used to address the problem of imbalanced data.
- 3) The train and test datasets are used to conduct the experiments on the proposed model.

The structure of the paper is organized as follows: The related work of existing algorithms is described in section II, while section III refers proposed hybrid model for fraud detection. Experimental credit card fraud detection, results, and discussion are presented in Section IV. The paper's conclusion is discussed in Section V in the final part.

## II. RELATED WORK

The performance of machine learning and data mining to prevent credit card fraud has been examined by the authors in [1]. On the other hand, most researchers used some classification measures to assess the solutions. A credit card detection model was used to extract the right attributes from transactional data. The aggregate approach was utilized to observe the customer's spending behavioral pattern. The author of this research proposes to construct a new set of features based on the periodic behaviors of transaction time

using an aggregation technique. An actual credit card fraud detection dataset from a large European cardholder company was used by the author. To examine the results, the author compared state-of-the-art credit card fraud models and weighed the pros and cons of various feature sets.

Credit cards are becoming more widely used in financial transactions, and at the same time, fraud is also increasing. The author presented a convolutional neural network framework to capture the pattern of fraud data in this research [2]. The author has proposed a trading entropy model to identify more complex consuming behaviors. Aside from that, the author merges the trending features into feature matrices for convolutional neural networks. As a result, the CNN model outperforms state-of-the-art approaches.

Supervised fraud classification algorithms for credit card fraud detection were proposed in [3]. The author has used two bank datasets to test these methods. Aggregation methods were suitable in many situations, but not all. SVM, logistic regression, random forest, and KNN were some of the classification algorithms used by the author. Out of this, the random forest gives better accuracy. Credit card transactions, as well as the fraud linked with them, are becoming more popular today. When credit card information is obtained unlawfully and used to make purchases, credit card fraud occurs. If credit card data is available and sufficient for a company or service, the author used a different machine learning technique to tackle the problem.

In [4], several popular methods in supervised, unsupervised, and ensemble classification were evaluated. The authors have applied different algorithms to identify fraudulent and legitimate transactions. Because unsupervised algorithms handle the skewness of datasets better than supervised algorithms, they outperform supervised algorithms in terms of performance measures. In future work, the author wants to contribute to the re-sampling techniques that will help us to balance data.

In [5], the authors have proposed a method to identify fraudulent and legitimate transactions. Because of the rapid progress of e-commerce and online banking, the usage of credit cards has increased dramatically, resulting in a large number of fraud instances. The author proposed a novel fraud detection method that has three stages. The first phase involves initial user authentication and card details verification. After the initial state, the transaction proceeds to the following step, where a fuzzy c-means clustering method was applied to determine the new pattern of credit card users based on their previous transactions. The authors used fuzzy c-means clustering algorithms to group similar datasets and a neural network to reduce misclassification based on the amount, timing, and kind of items purchased. For analyzing the proposed model, the author used stochastic models. The authors concluded that the application of fuzzy clustering and learning was the solution to a real-world problem based on the findings.

The main objective is to determine whether a transaction is legitimate or fraudulent. Various techniques, such as supervised and unsupervised procedures, were used to detect fraud [6]. Numerous methods identify fraud when utilizing

supervised techniques. The author combined supervised and unsupervised techniques to classify credit card fraud to build a hybrid approach to improve system accuracy. This based on the results using the hybrid model gives better accuracy.

In [7], the authors have proposed long short-term memory networks as a method to aggregate the new pattern of data purchase behavior of cardholders, to improve the accuracy of the credit card fraud system. The comparison of baseline random forest to long short-term memory in this research improves the detection of accuracy as offline transactions, where the cardholder was physically present at the merchant. The author looks at both sequential and non-sequential learning systems that benefit from aggregation strategies in this paper.

The authors have used different algorithms on real-time datasets such as nearest neighbors, random forest, naive Bayes, multiple Perceptron, ad boost, quadrant discriminative analysis, pipelining, and ensemble learning [8]. The sample consists of European cardholders who were present for two days in September 2013. The dataset is highly unbalanced, so the ADASYN method has been used to correct it. As for performance measures, the author used precision, recall, accuracy, F1-measure, Matthew's correlation coefficient, and Balanced Classification Rate. Depending on a variety of parameters the pipelining gives better accuracy.

The authors have proposed Fraud-BNC, a customized Bayesian Network Classifier (BNC) algorithm on a real-time credit card fraud dataset in [9]. The Hyper Heuristic Evolutionary Algorithm was used to create BNC automatically (HHEA). A categorization dataset provided by Pag Seguro, a well-known Brazilian online payment provider, caused this difficulty. The author deals with two issues: a skewed dataset and misclassified fraud costs. As a result of Fraud-BNC, the method's economic efficiency was evaluated and tested against seven alternative classification algorithms. When it comes to accuracy, Fraud-BNC outperforms other algorithms.

In business and banking, credit card fraud has become an issue. Credit card fraud occurs when a fraudster employs modern techniques and technology to complete credit card information without the owner's permission. The author proposed an intelligent approach for detecting credit card fraud using an upgraded light gradient boosting machine to address this issue (OLightGBM) in [10]. The author goes through numerous stages to establish this framework, including data collection, data pre-processing, model development, and model evaluation. The researcher had to use a Bayesian-based hyperparameter to maximize the parameter in the suggested approach. To assess the performance of the intelligent technique, the author employed two real-time datasets for detecting credit card fraud transactions. The first dataset comes from credit card fraud transactions made by European Cash Holders in 2013. The second dataset came from the UCSD-FICO Data Mining Contest in 2009. To compare with the provided technique, the author used a variety of machine learning algorithms. As a result, OLightGBM exceeds confusion matrices, accuracy, precision, and recall, among other performance metrics.

In today's technology world and internet banking, credit card usage is rapidly increasing. Credit cards have become the most frequent payment method for online purchases. As a result, the number of cases of fraud increased. Stopping fraud was critical since it hurts the economy. To solve this problem, the author [11], has employed several techniques, including logistic regression (LR), Naive Bayes (NB), Support Vector Machines (SVM), decision tree (DT), and K-nearest neighbors (KNN), as well as random forest (RF). The author proposed a new deep learning architecture based on Spark to detect fraud. After that, the author compared the proposed deep learning architecture and the machine learning algorithm. The author used accuracy, precision, and recall performance metrics to classify fraudulent and legitimate transactions. As a result, random forest generates more precise outcomes.

In [12], the authors have proposed a novel fraud detection system that evaluates customers' previous transaction records and extracts pattern behavior. At the start, the authors used the clustering method to separate the cardholders into groups. To determine cardholder behavior, researchers employ the sliding window method to organize transactions. The dataset contained the European cardholder dataset. To balance the credit card fraud dataset, the author applied SMOTE techniques. Another option for dealing with unbalanced datasets is to employ the single class SVM. The authors used a variety of algorithms, both with and without statistical methodologies, to determine the dataset's correctness. Local Outlier factor, Isolation Forest, Support vector machine, logistic regression, decision tree, and random forest are some of the algorithms used. The main objective of this paper is to classify fraud and legitimate transactions.

In [13], has developed a deep learning and machine learning method to detect credit card fraud transactions. For developed a model author performed data pre-processing, normalization, and under-sampling techniques using a European cardholder imbalanced dataset. Then compare the machine learning methods such as Support Vector Machine and K-Nearest neighbors. After that, to train, the model artificial neural network was used by the author. As a result, the artificial neural network gives better accuracy.

In [14], the authors have focused on a new ensemble learning algorithm that combined bagging and boosting. As a result, detecting credit card fraud is a difficult task. The author proposed an ensemble hybrid model with the bagging and boosting method. The authors perform steps like pre-processing and feature engineering with ad-boost divide the data between train and test groups, classify the test data set using bagging-based ensemble classifiers like random forest and extra tree approaches, and generate results. The dataset UCSD-FICO was used as an input, and it is a severely unbalanced dataset. As a result, the author of the balancing data set utilized various strategies. The original feature space to the next feature space mapping approach was utilized in the first step, followed by the generation of the feature space. The next step is to use a tree-based classifier to solve the classification and regression problems. The author employed false negative and false positive rates, detection rates, and accuracy rates in the proposed model.

To detect credit card fraud, the authors [15] have used a different machine learning method. The results of a benchmark and a real-world dataset were compared by the author. A hybrid method combining ad boost and majority voting has also been developed by the researcher. The author compared the performance of a single model and a hybrid model on the same dataset. Researchers used naive Bayes, random forest, Decision tree, Neural Network, Linear Regression, Deep Learning, Logistic Regression, SVM, and Multilayer Perceptron as machine learning techniques. As a result, the methods were utilized to assess using ad boost and majority voting, which improved the accuracy with benchmark and real-time datasets.

In [16], a semi-supervised technique to detect credit card fraud, in which user profile clusters were created and used to construct classifiers. Users were profiled and grouped based on their patterns of conduct. Consumer segments were spread and further divided based on transaction factors such as volume, frequency, and distance. Random forest and XGBoost classifiers were trained on the total sample and compared to transaction-level classifiers in each cluster. This study finds that classifiers trained at the cluster level need not improve classifiers trained in the sample group in terms of overall weighted performance. The clustering method was used to identify groups of account holders. Moreover, some classifiers trained in specific groups do significantly better than the baseline, whereas classifiers learned in other groups do not perform as well. The optimum classifier for a given cluster varies by cluster and demonstrates the potential for new classifiers to perform well on groups that currently use underperforming models.

For credit card fraud detection, a Decision tree and random forest are used [17]. The author has used public data as sample data to test the model's efficiency. The finding was similar to a set of real-world credit card data obtained from a financial institution. Furthermore, some clutter was introduced to the data samples as a secondary check on the system's endurance. The study's methods were significant in that the first method created a tree against the user's activity, and frauds were detected using this tree. A user activity-based forest will be generated in a second way, and an attempt will be made to identify the suspect using this forest.

Artificial intelligence techniques were used to classify a fraudulent transaction as a routine transaction [18]. The author was to compare and contrast the results of several machine learning algorithms in detecting credit card fraud. The algorithm's rank and performance are of primary interest to the author. The model for identifying bad transactions in the e-commerce dataset was analyzed using the UCSD-FICO Data mining content 2009 dataset; Performance measures used by the author, such as classification accuracy and fraud detection rate.

To deal with anomalous transactions and develop a cardholder behavior model was proposed in [19]. To classify fraudulent and legitimate activities, the author used classification algorithms such as naive Bayes, Bayes Net, random forest, j48, libSVM, and MOLEM, as well as the Weka tool. Initially, the data was created and tested using the

random forest and j48 models. The author used a real-time dataset to test the efficacy, and random forests performed better.

In [20], the authors have proposed decision trees, random forests, and logistic regression as machine learning algorithms for fraud detection. The analytical model was put to the test using the benchmark dataset. The most accurate models are the random forest and decision tree. A confusion matrix was employed to assess accuracy Table I.

TABLE I. LITERATURE SURVEY

Sr.no	Title	Dataset	Methodology & tools	Advantages and limitation
1	Feature engineering strategies for credit card fraud detection.	European cardholder dataset	Decision Tree, Logistic Regression, Random Forest	The author proposed a method to improve the performance results of credit card fraud. The author has a problem with the system it takes a long time to make a decision.
2	Credit card fraud detection using convolutional neural networks.” In International conference on neural information processing	Real-time credit card dataset	Convolution Neural network cost Estimation method	The advantage of implementing a convolutional neural network is to capture neural patterns of fraud activity discovered from a labeled dataset. Because there are far fewer fraud transactions in real life than in non-fraud situations, the major drawback when implementing it is the issue of an unbalanced dataset.
3	Transaction aggregation as a strategy for credit card fraud detection	Bank A and Bank B	SVM, logistic regression, Random Forest and KNN, Cart,	The advantage of aggregation is data do not need to be properly classified, it may be more resistant to the impacts of population drift.
4	Performance Evaluation of Machine Learning Algorithms for Credit Card Fraud Detection	European cardholder dataset	NB, RF, KNN LR, XGBT, SVM ANN, DL	The main advantage of unsupervised algorithms performs better throughout all measures both in absolute terms and in comparison, to other approaches since they are better at handling the dataset skewness.
5	Credit Card Fraud Detection: A Hybrid Approach Using Fuzzy Clustering & Neural	Real-time credit card dataset	Fuzzy Clustering and Neural Network	In this paper, fuzzy clustering is used to decrease the misclassified rates of transactions and also find new patterns based on past transaction data. The authors further

	Network			used the different attributes to correctly classify the transactions.	
6	Combining Unsupervised and Supervised Learning in Credit Card Fraud Detection	Credit card fraud detection dataset		K-means clustering, ensemble learning,	In this paper main advantage is a combination of supervised techniques and unsupervised techniques to improve accuracy.
7	Sequence Classification for Credit-Card Fraud Detection	Real-world fraud detection dataset.		Random Forest, long short-term memory	The advantage of implementing a neural network to identify credit card fraud is that it can identify credit card activity and use patterns in a significant amount of customer and transactional data.
8	Credit Card Fraud Detection using Pipeline and Ensemble Learning	European cardholder dataset		Logistic Regression, Naive Bayes, K nearest neighbors, Multi-Layer Perceptron, Ada Boost, Quadrant Discriminant Analysis, Random Forests, Pipelining, and Ensemble Learning	The author compared different algorithms in which pipelining works best as compared to another algorithm. The benchmark dataset was highly imbalanced so the author was able to the balanced dataset.
9	A customized classification algorithm for credit card fraud detection	Pag Seguro dataset		customized Bayesian Network Classifier (BNC) algorithm for a real credit card fraud detection problem	In this paper, the authors were given High processing and detection speed on the Pag Seguro dataset for credit card fraud detection.
10	An Intelligent Approach to Credit Card Fraud Detection Using an Optimized Light Gradient Boosting Machine	Real-world credit transaction		Optimized light gradient boosting machine (OLightGBM), light gradient boosting machine (LightGBM), KNN, SVM, NB, DT.	In this paper, OLightGBM gives better accuracy than other machine learning algorithms. The proposed model identifies a useful pattern of credit card fraud.
11	An Enhanced Secure Deep Learning Algorithm for Fraud Detection in Wireless Communication	European cardholders		Logistic regression (LR), Naive Bayes (NB), Support Vector Machines (SVM), decision tree (DT), and K-nearest	The credit card fraud dataset is highly imbalanced but in this system that imbalanced problem was solved. Credit card fraud prevention was a very important task but the author was unable to achieve it.



			neighbor (KNN) as well as random forest (RF)	
12	Credit card fraud detection using machine learning.	European cardholder dataset	DT, Local Outlier factor, Isolation forest, LR, RF	The advantage is that the author balanced the dataset using SMOTE techniques. In this paper, the authors need a balanced dataset for achieving high precision and recall.
13	Credit card fraud detection using artificial neural network	European cardholder dataset	SVM, KNN, and ANN	Using an artificial neural network model turns out to be the best for detecting fraud. The author was also unable to balance data using normalizing, under-sampling, or pre-processing methods.
14	Credit Card Fraud Detection by Modelling Behavior Pattern using Hybrid Ensemble Model	Brazilian bank data and UCSD-FICO data.	Ensemble learning techniques such as boosting and bagging.	Combination of Bagging and boosting ensemble learning method for distributing credit card detection. Dataset is highly imbalanced. For analyzing the behavior of the customer drift method
15	Credit card fraud detection using AdaBoost and majority voting	Benchmark Dataset. Real-time dataset	AdaBoost and majority voting methods	In this paper author used Majority and AdaBoost. The majority voting gives better accuracy. The author wants to use online learning methods where online learning methods prevent fraud it informs before fraud happens.
16	Improving Credit Card Fraud Detection by Profiling and Clustering Accounts	Credit card bank dataset	Random forest and XGBoost, k-mean clustering,	In this paper, k means clustering used for the effective and fast result of data. The disadvantage of k means is selected k values. The authors used Clustering to improve the detection of credit card fraud
17	A Novel Approach for Credit Card Fraud Detection using Decision Tree and Random Forest Algorithms	Credit card fraud Dataset	Bayesian network, Gaussian network, Random Forest, Decision Tree	The advantages of decision trees and random forests where random forests are utilized for preventing overfitting problems. The disadvantage of a decision tree, it gives the problem of overfitting.

18	An Evaluation of Computational Intelligence in Credit Card Fraud Detection	UCSD-FICO Data Mining Contest 2009 dataset	This paper analyses and compares various popular classifier algorithms that have been most commonly used in detecting credit card fraud	Classification accuracy and fraud detection rate are high in an evaluation of computational intelligence in credit card fraud detection. For credit card, and anomaly detection author want to propose a reliable expert system.
19	Credit Card Fraud Detection Based on Transaction Behavior	Real-time dataset	Random Tree and J48	Random forest gives the highest accuracy on the real-time dataset. The author had been unable to solve a random forest problem due to the slowness of the algorithm in a large number of trees.
20	Predictive Modelling for Credit Card Fraud Detection Using Data Analytics	German credit card fraud dataset	Logistic Regression, Decision Tree, Random Forest, Decision Tree.	The author compares various algorithms and concludes that random forest gives better accuracy. The author occurred a problem during the testing of random forest speed during the predictive model.

### III. PROPOSED HYBRID MODEL FOR FRAUD DETECTION

Fig. 1 shows the primary steps of the proposed models, which include data preparation, EDA, and a hybrid ensemble model. For the credit card fraud system, the result has been given as a categorization of genuine and fraudulent transactions. Data preparation employs the Smote technique, outlier removal, and null value deletion. After preprocessing, the hybrid ensemble model has been proposed for differentiating between legitimate and fraudulent transactions. A hybrid model reduces the risk of fraudulent transactions compared with a single model before applying SMOTE.

#### A. Data Preprocessing

Data pre-processing is an essential step because without it, the model can generate inaccurate results and it helps to preserve the integrity of the data. Data distribution, outlier identification, and noise reduction have been part of the data preprocessing stage.

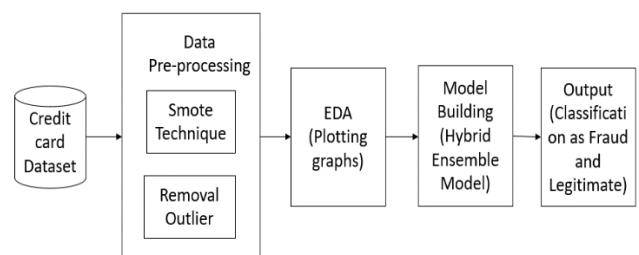


Fig. 1. Proposed Model of Credit Card Fraud Detection.

Anomaly detection is a method for finding outliers or odd patterns that deviate from anticipated behavior. In the proposed model Interquartile range (IQR) has been used to remove the outliers [21]. The interquartile range, or IQR, is the space between the first and third quartiles, or Q1 and Q3:

$IQR = Q3 - Q1$ . Outliers are data points that are either below or above the median ( $Q1 - 1.5 IQR$  or  $Q3 + 1.5 IQR$ ).

Q1 is the median.

Q2 is the average of the n smallest data points.

Q3 is the average of the n highest data points.

### B. Bagging-Based Ensemble Learning

Bootstrap aggregation, sometimes known as "bagging," is a common strategy applied in ensemble learning-based models that integrate both classification and regression techniques, hence increasing accuracy and other associated metrics. The principle of bagging is the combination of weak learners with a strong learner [14]. For our experimentation, we implemented decision tree-based bagging classifiers such as random forest-based classifiers.

In bootstrap sampling, replacement sampling is used to produce a bootstrap sample  $B_{Si}$  that is equal to  $D$ , where  $D$  is the input data. When  $D$  is big enough,  $B_{Si}$  acts as an independent version of  $D$ , and the assumed empirical distributions resemble  $D$  [15]. Therefore,  $B_{Si}$  might be viewed as a distinct and comparable variant of  $D$ . At bagging, in the  $i_{th}$  iteration, the model's predictions are averaged to suit the bootstrap sample  $B_{Si}$ .

In the end, bootstrap sampling wants to remove a classifier's potential for overfitting.

1) *Random forest*: A supervised machine learning approach based on ensemble learning is known as a random forest. To create a more efficient prediction model, you can combine several algorithms or use the same technique more than once in ensemble learning [3, 4]. The term "Random Forest" comes from the fact that the random forest method mixes several algorithms of the same type or different decision trees into a forest of trees. Both regression and classification tasks may be performed using the random forest approach.

The basic steps of the random forest method are as follow [21]:

- a) Choose  $N$  records at random from the dataset.
- b) Based on these  $N$  records, construct a decision tree.
- c) Repeat steps 1 and 2 after choosing how many trees you.
- d) Want in your algorithm.
- e) Each tree in the forest can forecast the category to which the new record belongs in a classification issue. The category that receives the majority of the votes is finally given a new record.

2) *Adaboost algorithm*: Adaboost is one boosting technique, which is similar to Random Forest Classifier. The

Ada-boost classifier combines weak and strong classifier algorithms to create a large classifier [8]. A single algorithm may incorrectly categorize the items; however, by combining many classifiers, selecting the training set at each iteration, and assigning the appropriate amount of weight in the final vote, we can achieve a high accuracy score for the entire classifier. It keeps the algorithm repeatedly by selecting the training set depending on prior training accuracy. At any iteration, the weighting of each trained classifier is determined by the accuracy achieved. Adaboost provides weight to each training item after training a classifier at any level. The weight of a misclassified item is increased so that it is more likely to appear in the training subset of the Adaboost classifier. The basic steps of the Adaboost algorithm [21] are:

a) Initialize  $M$ , the maximum number of models to be fit, and set the iteration counter  $m=1$ .

b) Initialize the observation weights  $w_i = 1/N$  for  $i = 1, 2, N$ . Initialize the ensemble model  $\hat{f}_b = 0$ .

c) Train a model using  $\hat{f}_m$  observation weights that minimize the weighted error  $ECM$  defined by summing the weights for the misclassified observations is shown in (1).

d) Add the model to the ensemble:

$$\hat{f}_m = (\hat{f}_{m-1}) + (\hat{\alpha}_m)(\hat{f}_m) \quad (1)$$

$$\text{Where } (\hat{\alpha}_m) = \frac{\log(1-e)m}{e_m} \quad (2)$$

e) Update the weights  $w_1, w_2, w_3, \dots, w_N$  so that the weights are increased for the observations that were misclassified. The size of the increase depends on  $(\hat{\alpha}_m)$  with larger values  $(\hat{\alpha}_m)$  leading to bigger weights as mentioned in (2).

f) Increment the model counter  $m=m+1$  if  $m_i=M$ , go to step 1. The boosted estimate is given below:

$$\hat{f} = (\hat{\alpha}_1) (\hat{f}_1) + (\hat{\alpha}_2) (\hat{f}_2) + \dots + (\hat{\alpha}_m) (\hat{f}_m) \quad (3)$$

g) The factor  $(\hat{\alpha}_m)$  has a lower error and higher weight.

### C. Logistic Regression

One of the most widely used Machine Learning algorithms, within the category of Supervised Learning, is logistic regression. The categorical dependent variable is predicted using a collection of independent factors. In a categorical dependent variable, the output is predicted using logistic regression [3]. As a result, the result must be a discrete or classifying value. Instead of providing the exact values of 0 and 1, it gives the probabilistic values that fall between 0 and 1. It can be either Yes or No, 0 or 1, true or false. Except for how they are applied, logistic regression and linear regression are very similar [4]. While logistic regression is used to solve classification difficulties, linear regression is used to solve regression problems. In logistic regression, we fit an "S" shaped logistic function, which predicts two maximum values, rather than a regression line (0 or 1) is shown in Fig. 2. The logistic function's curve shows the possibility of several things, like whether or not the cells are malignant, and whether or not a rat is fat depending on its weight [11]. Because it can classify new data using both continuous and

discrete datasets, logistic regression is a significant machine learning approach.

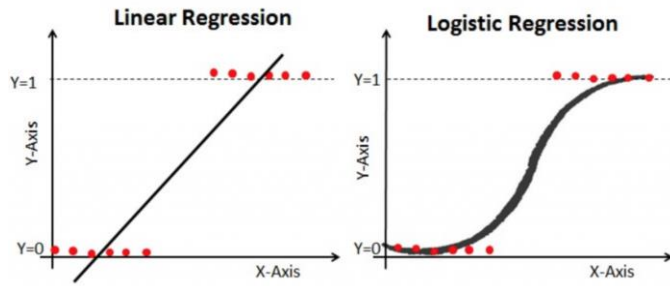


Fig. 2. Comparison between Linear Regression and Logistic Regression.

The equation of logistic regression of straight line written as [22]:

$$y = a_0 + a_1 * x_1 + a_2 * x_2 + \dots + a_k * x_k \quad (4)$$

In logistic regression, y can be between 0 and 1 only, so divide the above equation by (y-1):

$$\frac{y}{y-1} | 0 \text{ for } y = 0 \text{ and } \infty \text{ for } y = 1 \quad (5)$$

As a result, the logistic regression equation is defined as:

$$\log \frac{y}{y-1} = a_0 + a_1 * x_1 + a_2 * x_2 + \dots + a_k * x_k \quad (6)$$

#### D. Hybrid Ensemble Model

Fig. 3 shows how the hybrid ensemble model works. In a hybrid ensemble model, the first step is to train the model, and once it has generated individual results, the hybrid ensemble combines those outcomes with the help of majority voting to produce the final predicted results. The hybrid ensemble model is a combination of the bagging and boosting models. There are two well-known types of ensemble learning; bagging and boosting [14]. The widely used ensemble learning model for bagging is a random forest. Another well-liked ensemble learning approach that comes under the boosting category is AdaBoost. While the boosting models use the complete dataset, the bagging models only use a portion of the datasets. Random forest and Adaboost are employed as weak learners to create a hybrid ensemble model.

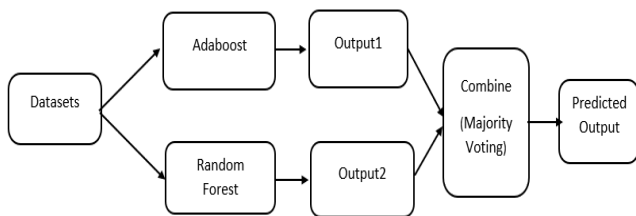


Fig. 3. Hybrid Ensemble Model.

#### IV. EXPERIMENTAL DETECTION OF CREDIT CARD FRAUD DETECTION

This section explains the specificity and stability of our proposed models and compares them with the most recent research-based models. Our main objective is to increase the model's capacity for fraud detection. A better understanding of the data is required to do this. The experimental study was conducted on a simple Windows computer with a quad-core processor and 8 GB of RAM, and the results on the European credit card dataset and the Credit card stimulation dataset were acceptable. The proposed system has implemented the hybrid ensemble model for classification of the credit card fraud detection using python programming on a Jupyter Notebook. This system, primarily apply smote technique for data imbalanced problem. Further implementation of the hybrid ensemble model is on credit card fraud dataset. For checking the performance of the model, precision, recall, F1- score, and ROCAUC are calculated for every test case.

##### A. Data Description

Table II shows the instances, columns, and fraudulent and non-fraudulent cases of the European credit card dataset and credit card fraud stimulation dataset. Datasets are used to train and validate the efficacy of proposed approaches and hence play an essential part in research motivation. In this section, we'll go through two different datasets that have been used in our suggested approach's experiments.

1) *European dataset:* The first dataset, collected from www.kaggle.com, consists of credit card transactions performed by European cardholders within two days in September 2013, with 492 fraudulent transactions out of 284,807 as shown in Fig. 4. It has 31 features, including the time when a transaction occurred, the number of transactions, and 28 other qualities labeled V1 to V28, as well as the target label 'Class,' which uses a binary value of '1' or '0' to determine if a transaction is fraudulent or not [13].

TABLE II. THE CREDIT CARD DATASET DESCRIPTION

Name	Instances	Features	Normal	Fraudulent
European Credit Card Dataset	284,807	31	248,315	492
Credit Card Stimulation Dataset	594,643	10	587,443	7200

V6	V7	V8	V9	...	V21	V22	V23	V24	V25	V26	V27	V28	Amount	Class
0.462388	0.229599	0.090898	0.363787	...	-0.018307	0.277838	-0.110474	0.068828	0.128538	-0.189115	0.133588	-0.021053	149.62	0
-0.082361	-0.078803	0.085102	-0.255425	...	-0.225775	-0.638672	0.101288	-0.338046	0.167170	0.125885	-0.008893	0.014724	2.69	0
1.800499	0.791461	0.247676	-1.514654	...	0.247998	0.771679	0.909412	-0.688281	-0.327842	-0.139097	-0.055353	-0.059752	378.66	0
1.247203	0.237609	0.377436	-1.387024	...	-0.108300	0.095274	-0.190321	-1.175575	0.647376	-0.221829	0.062723	0.061458	123.50	0
0.095921	0.592941	-0.270533	0.817739	...	-0.008431	0.798278	-0.137458	0.141267	-0.206010	0.502292	0.219422	0.215153	69.99	0

Fig. 4. European Credit Card Dataset.

2) *Credit card bank stimulation dataset*: The second dataset includes 594,643 transactions made across 180 simulated days, 7200 of which are considered fraudulent (1.2 percent). This dataset is a synthetic dataset developed with BankSim software, which is a simulation tool meant to simulate fraud data in Fig. 5, BankSim uses a multi-agent simulation methodology that is based on a sample of aggregated real-time transaction data provided by a Spanish bank. Thousands of transactional data records from November 2012 to April 2013 make up the initial bank data. To simulate this genuine bank data, BankSim employs many agents from three main categories: traders, customers, and fraudsters. These agents interact with one another for a duration of a few days, establishing a purchasing transaction log that strongly matches the original bank.

step	customer	age	gender	zipcodeOri	merchant	zipMerchant	category	amount	fraud
0	'C1093826151'	'4'	'M'	'28007'	'M348934600'	'28007'	'es_transportation'	4.55	0
1	'C352968107'	'2'	'M'	'28007'	'M348934600'	'28007'	'es_transportation'	39.68	0
2	'C2054744914'	'4'	'F'	'28007'	'M1823072687'	'28007'	'es_transportation'	26.89	0
3	'C1760612790'	'3'	'M'	'28007'	'M348934600'	'28007'	'es_transportation'	17.25	0
4	'C757503768'	'5'	'M'	'28007'	'M348934600'	'28007'	'es_transportation'	35.72	0

Fig. 5. Credit Card Bank Stimulation Dataset.

**B. Performance Parameter**

The learning algorithm's performance measure showed unbalanced behavior in the imbalanced distribution of the classes. So, it is necessary to select suitable measures to assess the effectiveness of the categorization system. Precision, recall, F1 Score, and accuracy have been chosen as performance evaluation metrics for the proposed work because the learning algorithm exhibits an accuracy phenomenon in unbalanced scenarios.

True Positive (TP) - How many safe cases did our model properly predict.

False Negative (FN) - How many cases of our model incorrectly predicted.

False Positive (FP) - How many fraud cases are classified incorrectly.

True Negative (TN) - How many fraud cases are classified correctly.

Precision -  $TP / (TP + FP)$

Recall -  $TP / (TP + FN)$

F1 Score - Harmonal mean of precision and recall

F1 Score =  $(2 * precision * recall) / (precision + recall)$

**C. Data Imbalanced**

In this paper, two benchmark datasets have been used. In the given datasets there are few fraudulent incidence which makes data imbalanced. The Fig. 6 and Fig. 7, represents percentage ratio of fraudulent transaction of the European credit card dataset and the credit card stimulation dataset

respectively. To increase the fraudulent cases Synthetic Minority Oversampling Technique (SMOTE) technique is used. The Fig. 8, represents the number of increases the fraudulent cases after applying smote approach for the European credit card dataset and the Credit card stimulation dataset.

**D. Exploratory Data Analysis**

Exploratory data analysis is a data analysis process used to properly understand the data and discover its many characteristics, frequently using visual methods. Data analysis allows us to better understand and identify meaningful patterns in it.

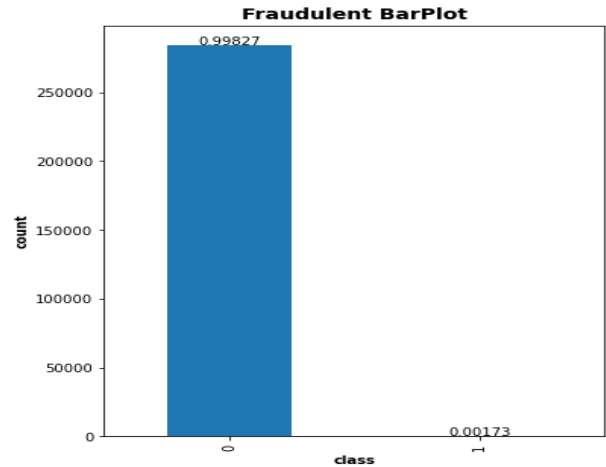


Fig. 6. European Credit Card Dataset before Applying Smote Technique.

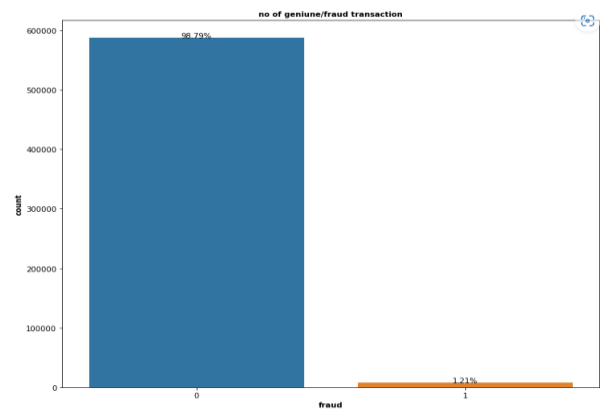


Fig. 7. Credit Card Stimulation Dataset before Applying Smote Technique.

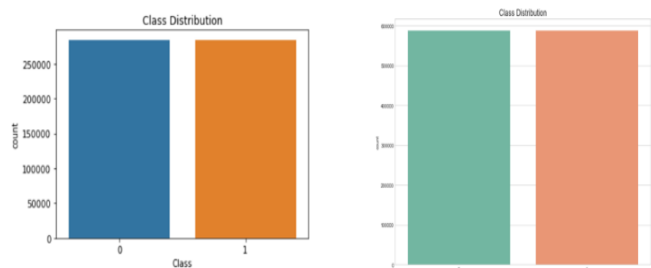


Fig. 8. European Credit card and Credit Card Stimulation Dataset after Smote Technique.

To analyze the time and amount in this paper, exploratory analysis is performed. Fig. 9 and 10, show how the time and number of transactions during the day and at night differ.

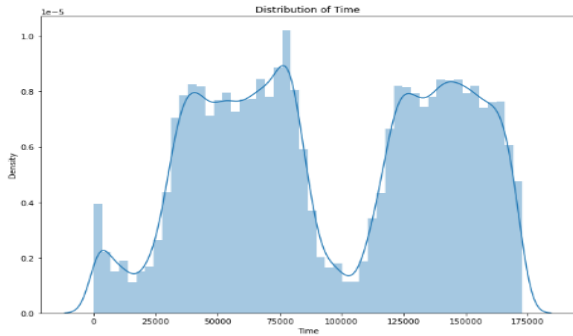


Fig. 9. Transaction Time of European Cardholder Dataset.

Fig. 9 shows the low peak value in the time distribution because there is a significant difference between the night and day transactions. In the density plot to x-axis shows the time of transaction and the y-axis shows the density of attributes.

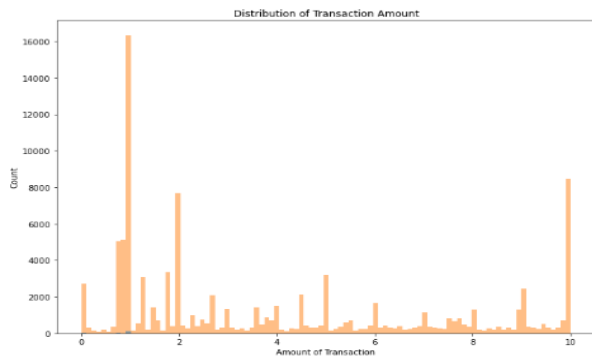


Fig. 10. Transaction Amount of European Cardholder Dataset.

Fig. 10 represents the total amount of money transacted. The majority of transactions are small, and just a few come close to reaching the maximum transaction value.

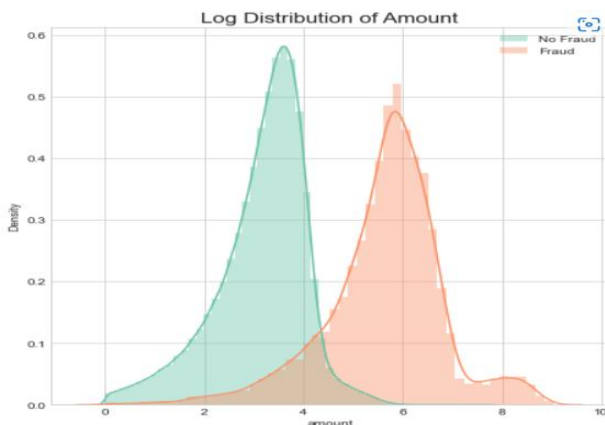


Fig. 11. Transaction Amount of Credit Card Stimulation Dataset.

Fig. 11 represents the total amount of money transacted. It shows that number of fraudulent transactions is less than a non-fraudulent transaction.

### E. Result and Discussion

Hybrid ensemble modeling is proposed to categorize fraudulent and legal transactions. The experiment is done on the European credit card and credit card stimulation dataset. The 70:30 % ratio is used for training and testing classifier. In both, the dataset fraudulent instances are less compared to non-fraudulent transactions. So, this is a serious issue that has been found with the dataset. In the European dataset 495 fraudulent transactions out of 284,807 non-fraudulent transactions, so the number of transactions needs to be increased. Same as credit card stimulation dataset 7200 fraudulent transactions out of 594,643. Our approach provides non-fraudulent transactions more weight when applied to an imbalanced dataset. Ensemble models are used to solve the main issue in credit card fraud detection, which is predicting future transaction behavior and finding the right solution.

The initial comparison between the single model and the original dataset is carried out in this study. But the single model has obtained less True positive value and more false positive value which indicates that more fraudulent transactions are presented in datasets because of unbalanced datasets. In order increases true positive value and handle unbalanced dataset oversampling strategies is used. To increases performance of the model hybrid ensemble model is proposed. The hybrid ensemble model is constructed by combining an Adaboost and random forest. This will improve the performance parameters of the system.

In this paper, the performance of the proposed hybrid ensemble model is compared to the machine learning algorithms, including logistic regression, random forest, and Adaboost. Table III and Table V shows the performance measure of European credit card and credit card bank stimulation on the imbalanced dataset. As we observed that the single model and ensemble model did not improve the true positive rate and true negative rate with the imbalanced dataset. So, we applied smote technique to the balanced dataset. Smote technique is used to increase fraudulent instances. After applying Smote oversampling method, the datasets are much more balanced.

The balanced dataset is added and tested with an ensemble model such as bagging and boosting and a predictive model such as logistic regression. Table IV is showing an improvement in precision, recall, and F1 score for the European dataset. Same for Credit card fraud detection dataset precision, recall, and F1 score slightly increased is observed in Table VI precision-recall (AUPR) curve is being used to analyses the performance measure of the proposed model. Table VII shows the comparison results of the European cardholder dataset and Credit card stimulation dataset, which show the AUPR score for boosting, LR, and Adaboost (+) random forest. Fig. 12 and Fig. 13 show the AUPR curve for random forest + Adaboost. On the European cardholder dataset and Credit card stimulation dataset, the proposed method shows a Random Forest +Adaboost which means the hybrid ensemble model gives better results than the single model.



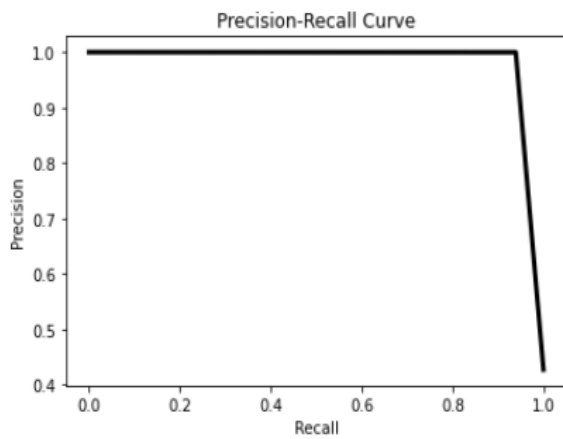


Fig. 12. Recall and Precision Curve for European Credit Card Dataset.

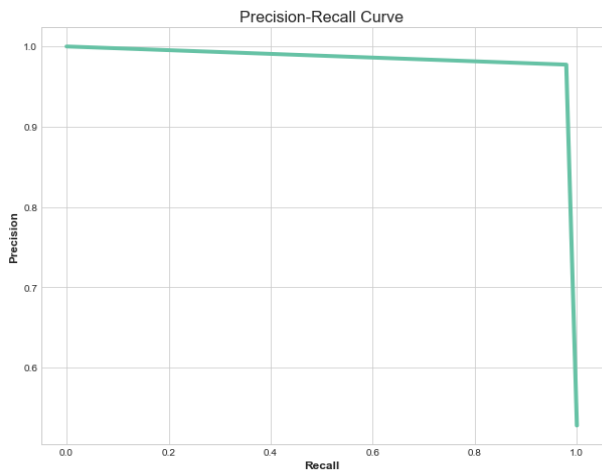


Fig. 13. Recall and Precision Curve for Credit Card Stimulation Dataset.

Table III and Table V shows the performance measure of classification algorithms on the European dataset and Credit card stimulation dataset before applying smote technique. So, we observed that after applying smote technique value of precision, recall, and F1 Score is improved rather to the without applying smote technique. Here hybrid ensemble model with random forest and Adaboost gives better precision, recall, and F1 Score.

Table IV and Table VI show the performance of measure of classification algorithm on European cardholder dataset and credit card stimulation dataset after applying smote technique. Here, the hybrid ensemble model with a combination of random forest and Adaboost gives better precision, recall, and F1- score than other algorithms.

TABLE III. BEFORE SMOTE TECHNIQUE ON EUROPEAN CREDIT CARD FRAUD DATASET

Algorithms	Precision	Recall	F1 Score
Logistic Regression	0.88	0.62	0.73
Adaboost	0.78	0.66	0.72
Random Forest	0.94	0.77	0.85
Random Forest+Adaboost	0.94	0.78	0.85

TABLE IV. AFTER SMOTE TECHNIQUE ON EUROPEAN CREDIT CARD FRAUD DATASET

Algorithms	Precision	Recall	F1 Score
Logistic Regression	0.96	0.90	0.93
Adaboost	0.97	0.94	0.95
Random Forest	0.97	0.98	0.95
Random Forest+Adaboost	1.00	0.94	0.97

TABLE V. BEFORE SMOTE TECHNIQUE ON CREDIT CARD FRAUD STIMULATION DATASET

Algorithms	Precision	Recall	F1 Score
Logistic Regression	0.88	0.62	0.73
Adaboost	0.78	0.66	0.72
Random Forest	0.94	0.77	0.85
Random Forest+Adaboost	0.91	0.78	0.84

TABLE VI. AFTER SMOTE TECHNIQUE ON CREDIT CARD FRAUD STIMULATION DATASET

Algorithms	Precision	Recall	F1 Score
Logistic Regression	0.92	0.99	0.97
Adaboost	0.97	0.99	0.98
Random Forest	0.98	0.97	0.98
Random Forest+Adaboost	0.99	0.99	0.99

The Predictive behavior of the proposed model is analyzed concerning the area under precision and recall curve. The result shown in Table VII is the area under precision and recall score for LR, boosting Adaboost +random forest. Here we observed that a hybrid ensemble model with random forest +Adaboost gives a better AUC Score than other algorithms.

TABLE VII. AREA UNDER CURVE SCORE ON EUROPEAN CREDIT CARD DATASET AND CREDIT CARD STIMULATION

The area under curve score (AUC Score)	European credit card dataset	Credit card bank stimulation dataset
Logistic Regression	95.10	95.80
Adaboost	96.79	98.43
Random Forest	97.32	98.09
Random Forest+Adaboost	98.26	99.37

## V. CONCLUSION

In this paper, a hybrid ensemble-based model is proposed to classify fraudulent and legitimate transactions. At the beginning of the project, the data analysis technique is used to map the original feature. To overcome the imbalanced problem oversampling smote method is used to balance the dataset during data pre-processing. After pre-processing, logistic regression, random forest, and Adaboost are used to check whether a transaction is legitimate or fraudulent. The hybrid ensemble model before applying the smote technique gives 0.85 % F1-Score and after applying smote technique it gives 0.97% on the European credit card fraud dataset. So, the F1 -score of Smote technique with the hybrid ensemble model



gives more results. For the Credit card stimulation dataset before applying smote technique F1 score gives 0.84% and after applying smote technique F1 Score gives 0.99%. It is observed that a hybrid ensemble model that combines random forest and Adaboost gives better results. From Table VI, the hybrid ensemble model for the European dataset achieves 98.27 % area under the curve score, whereas the Credit card fraud stimulation dataset achieves 99.37 % area under the curve score. In future work, apply the under-sampling method to check the performance of the algorithm and also use deep learning techniques for the classification of fraudulent and legitimate transactions.

#### REFERENCES

- [1] Bahnsen, Alejandro Correa, Djamilia Aouada, Aleksandar Stojanovic, and Björn Ottersten." Feature engineering strategies for credit card fraud detection." *Expert Systems with Applications* 51 (2016): 134-142.
- [2] Fu, Kang, Dawei Cheng, Yi Tu, and Liqing Zhang." Credit card fraud detection using convolutional neural networks." In *International conference on neural information processing*, pp. 483-490. Springer, Cham, 2016.
- [3] Whitrow, Christopher, David J. Hand, Piotr Juszczak, David Weston, and Niall M. Adams." Transaction aggregation as a strategy for credit card fraud detection." *Data mining and knowledge discovery* 18, no. 1 (2009): 30-55.
- [4] Mittal, Sangeeta, and Shivani Tyagi." Performance evaluation of machine learning algorithms for credit card fraud detection." In *2019 9th International Conference on Cloud Computing, Data Science Engineering (Confluence)*, pp. 320-324. IEEE, 2019.
- [5] Behera, Tanmay Kumar, and Suvasini Panigrahi." Credit card fraud detection: a hybrid approach using fuzzy clustering neural network." In *2015 second international conference on advances in computing and communication engineering*, pp. 494-499. IEEE, 2015.
- [6] Carcillo, Fabrizio, Yann-A`el Le Borgne, Olivier Caelen, Yacine Kessaci, Fr`ed`eric Obl`e, and Gianluca Bontempi." Combining unsupervised and supervised learning in credit card fraud detection." *Information sciences* 557 (2021): 317-331.
- [7] Jurgovsky, Johannes, Michael Granitzer, Konstantin Ziegler, Sylvie Calabretto, Pierre-Edouard Portier, Liyun He-Guelton, and Olivier Caelen." Sequence classification for credit card fraud detection." *Expert Systems with Applications* 100 (2018): 234-245.
- [8] Bagga, Siddhant, Anish Goyal, Namita Gupta, and Arvind Goyal." Credit card fraud detection using pipeline and ensemble learning." *Procedia Computer Science* 173 (2020): 104-112.
- [9] de S`a, Alex GC, Adriano CM Pereira, and Gisele L. Pappa." A customized classification algorithm for credit card fraud detection." *Engineering Applications of Artificial Intelligence* 72 (2018): 21-29.
- [10] Taha, Altyeb Altaher, and Sharaf Jameel Malebary." An intelligent approach to credit card fraud detection using an optimized light gradient boosting machine." *IEEE Access* 8 (2020): 25579-25587.
- [11] Sanober, Sumaya, Izhar Alam, Sagar Pande, Farrukh Arslan, Kantilal Pitambar Rane, Bhupesh Kumar Singh, Aditya Khamparia, and Mohammad Shabaz." An enhanced secure deep learning algorithm for fraud detection in wireless communication." *Wireless Communications and Mobile Computing* 2021 (2021).
- [12] Sailusha, Ruttala, V. Gnaneswar, R. Ramesh, and G. Ramakoteswara Rao." Credit card fraud detection using machine learning." In *2020 4th international conference on intelligent computing and control systems (ICICCS)*, pp. 1264-1270. IEEE, 2020.
- [13] Asha, R. B., and Suresh Kumar KR." Credit card fraud detection using artificial neural network." *Global Transitions Proceedings* 2, no. 1 (2021): 35-41.
- [14] Karthik, V. S. S., Abinash Mishra, and U. Srinivasulu Reddy." Credit Card Fraud Detection by Modelling Behaviour Pattern using Hybrid Ensemble Model." *Arabian Journal for Science and Engineering* 47, no. 2 (2022): 1987-1997.
- [15] Randhawa, Kuldeep, Chu Kiong Loo, Manjeevan Seera, Chee Peng Lim, and Asoke K. Nandi." Credit card fraud detection using AdaBoost and majority voting." *IEEE Access* 6 (2018): 14277-14284.
- [16] Kasa, Navin, Andrew Dabhura, Charishma Ravoori, and Stephen Adams." Improving credit card fraud detection by profiling and clustering accounts." In *2019 Systems and Information Engineering Design Symposium (SIEDS)*, pp. 1-6. IEEE, 2019.
- [17] Dileep, M. R., A. V. Navaneeth, and M. Abhishek." A novel approach for credit card fraud detection using decision tree and random forest algorithms." In *2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV)*, pp. 1025-1028. IEEE, 2021.
- [18] Mahmud, Mohammad Sultan, Phayung Meesad, and Sunantha Sodsee." An evaluation of computational intelligence in credit card fraud detection." In *2016 International Computer Science and Engineering Conference (ICSEC)*, pp. 1-6. IEEE, 2016.
- [19] Kho, John Richard D., and Larry A. Vea." Credit card fraud detection based on transaction behavior." In *TENCON 2017-2017 IEEE Region 10 Conference*, pp. 1880-884. IEEE, 2017.
- [20] Patil, Suraj, Varsha Nemade, and Piyush Kumar Soni." Predictive modeling for credit card fraud detection using data analytics." *Procedia computer science* 132 (2018): 385-395.
- [21] Peter Bruce, Andrew Bruce, Peter Gedeck." *Practical statistical for data scientists*. O'Reilly Media, 2017.
- [22] Alenzi, Hala Z., and Nojood O. Aljehane." Fraud detection in credit cards using logistic regression." *International Journal of Advanced Computer Science and Applications* 11, no. 12 (2020).

# Covid-19 and Pneumonia Infection Detection from Chest X-Ray Images using U-Net, EfficientNetB1, XGBoost and Recursive Feature Elimination

Munindra Lunagaria, Vijay Katkar, Krunal Vaghela  
Department of Computer Engineering  
Marwadi University  
Rajkot, India

**Abstract**—The pandemic caused by the COVID-19 virus is the most serious current threat to the public's health. For the purpose of identifying patients with Covid-19, Chest X-Rays have proven to be an indispensable imaging modality for the hospital. Nevertheless, radiologists are needed to commit a significant amount of time to their interpretation. It is possible to diagnose and triage cases of Covid-19 effectively and rapidly with the assistance of precise computer systems that are powered by Machine Learning techniques. Machine Learning techniques such as Deep Feature Extraction can help detect the disease with improved precision and speed when used in conjunction with X-Ray images of the lung. This helps to alleviate the problem of lack of testing kits. Using the U-Net for Semantic image segmentation for lung segmentation and deep feature extraction-based strategy that was suggested in this research, it is possible to differentiate between patients who have contracted the Covid-19 virus, pneumonia and healthy people. XGBoost and recursive feature extraction based proposed methodology is evaluated using 20 different Pre-Trained deep learning based models including EfficientNet variations and it is observed that the maximum detection accuracy, precision, recall specificity, and F1-score are achieved when EfficientNetB1 is used to extract deep features. The respective values for these metrics are 97.6%, 0.964, 0.964, and 0.982. These findings lend credence to the efficiency of the proposed methodology.

**Keywords**—Covid-19; u-net; efficientnet; semantic image segmentation; XGboost; recursive feature extraction

## I. INTRODUCTION

The recently found Coronavirus pneumonia, which has been given the name Covid-19 ever since it was found, is both extremely contagious and highly pathogenic [1]. The symptoms may advance to a more unadorned form of pneumonia, which can lead to serious health problems and the failure of multiple organs. There is also the risk that pneumonia will prove fatal to the patient. Covid-19 has spread to 223 nations and is blamed for over 6.4 million deaths around the earth; additionally, quantity of deaths caused by virus is steadily rising [2] (see Fig. 1). In any region of the world, medical professionals, governments, organizations, and countries face a noteworthy impediment in the form of need to diagnose Covid-19 patients as early as possible.

Even though immunological tests are offered in a number of countries, the RT-PCR i.e., Real-Time Reverse Transcription-Polymerase Chain Reaction is the process that is employed majority of the time to identify Covid-19. Early treatment can be ensured by recognizing the detrimental effects of Covid-19 by assessing photographs of patients' lungs rather than relying on RT-PCR, which has limited sensitivity (60–70%) and is also a technology that requires a significant amount of time.

Care for patients with Covid-19 necessitates careful tracking of their condition's evolution over time. When used in conjunction with other diagnostic tests, medical imagery practices like Computed Tomography (CT) and X-Rays of Chest can help to confirm a diagnosis of Covid-19 pneumonia and keep tabs on how the disease is progressing. These photos demonstrate the rapid progression of ground-glass opacities with irregular patterns after the onset of Covid-19 symptoms [3].

The term “Artificial Intelligence” (AI) refers to an assorted methodologies intended at imitating human cognition and deeds. Algorithms that consent high end processors to comprehend intricate patterns and associations from empirical information are the focus of the discipline of Machine Learning (ML), a subfield of AI. While traditional ML approaches are limited in their ability to deal with complicated problems like medical image categorization, Deep Learning (DL) takes cues from biological neural networks to attain greater power and flexibility [4].

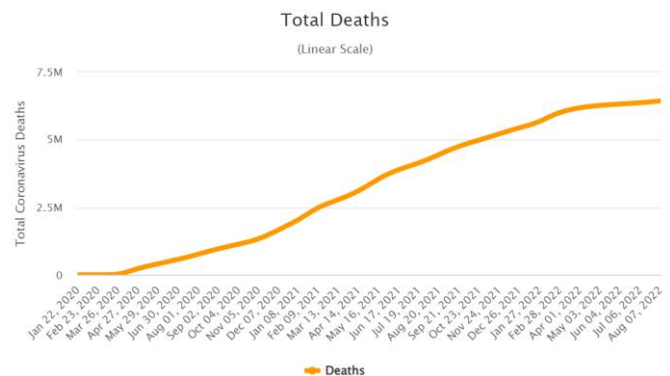


Fig. 1. Number of Worldwide Deaths Due to Covid-19.

Having access to a big dataset that contains images that have been labeled is essential to the success of a detection or classification system that is based on DL. Deep feature extraction (DFE) and transfer learning (TL) are two of the most successful alternatives to using a small sample size of images when there is not a big sample size of images available. TL refers to the process of making use of previously taught models to tackle unanticipated obstacles. TL is not an individual category of ML algorithms; rather, it is a process that can be utilized in the development of a new ML model. Model will be able to put the knowledge and abilities that it has gained in previous training to use in brand new situations [5]. In a manner analogous to the activity that came before it, this one will need organizing data according to the type of file. Another application of TL is the extraction of deep feature information. It is possible to extract feature vectors by making use of pre-trained CNN models rather than manually adjusting the activation layers of the CNN. The deeper layers, which are triggered by the activations of the lower-level layers, include the higher-level features that are essential to the classification of images [6].

Using U-Net, EfficientNetB1, XGBoost and Recursive Feature Elimination (RFE), this paper provides a DFE-based technique for discriminating healthy individuals and those infected with pneumonia from those infected with Covid-19. The remaining paper is structured as follows. The relevant work is briefly described in Section II, and use of U-Net model is explained in Section III. Brief introduction of Pre-Trained models employed in the study is provided in Section IV. The proposed approach is explained in Section V whereas dataset utilized for the experiment along with experimental findings is described in Section VI. Section VII briefly describes the significant contributions of the research work whereas section VIII concludes the research work.

## II. LITERATURE SURVEY

CNNs are constructed by first adding a sequence of layers that are known as convolutional, then adding layers that are known as normalization, and finally adding layers that are known as pooling. Feature extraction is the responsibility of the convolution layers, whereas feature normalization and feature down sampling are the purview of the normalization and pooling layers, respectively. In a manner analogous to that of more traditional approaches to machine learning, training of the CNN is carried out by use of an optimization strategy like Stochastic Gradient Descent [7].

A DL-based framework with fuzzy enhancements was presented by Cosimo Ieracitano et al. [8] under the name CovNNet for the purpose of classifying Chest X-Rays of Covid-19 patients and recognizing those of other patients with pneumonia. The CovNNet model was given X-Rays of the chest as well as fuzzy images that were created using a fuzzy logic based edge detection method. Experiments have shown that using fuzzy features in conjunction with chest X-Rays can result in enhanced classification performance.

Emtiaz Hussain et al. [1] have introduced a 22-layer deep CNN model named CoroDet with the purpose of being utilized in the process of recognizing Covid-19 cases in chest CT and X-Ray images. Their methodology is validated for its

usefulness utilizing a five-fold cross-testing procedure. Patient detection method using X-Ray pictures that is based on a faster RCNN-based object detection methodology was proposed by Fátima A. Saiz and Iñigo Barandiaran [9].

To develop a concatenated neural network, Mohammad Rahimzadeh and bolfazi Attar [10] merged features extracted with ResNet50V2 and Xception, then transmitted the aggregated data to a convolutional layer. The convolutional layer is responsible for the extraction of features, and it uses 1024 filters with a Kernel size of 1x1. This layer is introduced in order to provide assistance to the network in better learning characteristics from the combined data. They demonstrated the viability of their approach by utilizing a procedure known as three-way cross-testing.

Arman Haghanifar et al. [11] compiled a significant number of different chest X-Ray images from a variety of sources into a single extensive database that is open to the public. The TL paradigm is used in conjunction with the CheXNet model to generate a model that is referred to as COVID-CXNet. This model is effective at detecting pneumonia caused by the coronavirus and pinpointing its exact location.

Mohammad Shorfuzzaman et al. [12] describes an ensemble-based DFE approach as a means of determining the existence of Covid-19 in X-Ray of chest images. This technique takes the weights that were determined by a large number of models that have been pre-trained and turns them into a single band that represents X-ray characteristics. After that, these traits will be utilized in the process of making a diagnosis of the ailment.

M. D. Kamrul Hasan et al. [13] suggest a two-stage, DFE-based process that makes use of VGG16 Net and patient lung X-Ray photos in order to sense COVID-19-induced pneumonia. When evaluating their methodology, the scores for precision, specificity, and sensitivity are taken into consideration.

Junfeng Li et al. [14] developed COVID-GATNet with the intention of identifying Covid-19 patients with the assistance of Graph Attention Network. The results of the X-Ray input can be interpreted as normal, as having Covid-19 +, or as having pneumonia. They used the information from three different open-source datasets to produce a more extensive training data pool for their studies.

Debabrata Dansana et al. [15] proposed a three-stage approach by utilizing CNN, Pre-Trained CNN models, and Decision Tree. The first step in the process is to use CNN to clean up the input images by removing any noise that isn't wanted and building composite feature maps. The performance of the pre-trained CNN models that are used to extract image features is improved with the help of these Composite feature maps, which are used in the second step. Second stage extracts image features to use for training a Decision Tree, which will be used for categorizing the disease.

According to research that was carried out by Karim Hammoudi et al. [16] who optimized four different Pre-Trained CNN models (specifically ResNet50, InceptionResnetV4, Densenet, and VGG-19), InceptionResnetV2 had the lowest

false negative rate. Abdullahi Umar Ibrahim et al. [17] showed that optimizing the Alexnet Pre-Trained model for patient detection in X-Ray pictures led to the discovery that the model is capable of providing good detection precision. Asif Iqbal Khan et al. [18] developed a TL-based model utilising Xception Net and gave it the name CoroNet. Xception Net's base layers are connected to the dropout layer, which is in turn coupled to fully-connected layers for classification task. A DFE-based approach employing the VGG-19 Pre-Trained Model was presented by Harsh Panwar et al [19]. The original VGG-19 model's foundational layers were expanded with fully connected layers, and they were then fine-tuned for patient detection.

Eduardo Luz et al. [20] presented DFE based method of hierarchical classification with the help of EfficientNet models. Target categories are positioned at leaves of the tree, whereas classifiers are located in the in-between nodes of the tree. At the root node, one classifier was used to segregate between the Normal and Pneumonia patients, while at a higher level, another classifier was used to segregate between the Pneumonia patients themselves.

Two-stage DFE based model for Covid patient detection using U-Net and Pre-Trained models was presented by Sivaramakrishnan Rajaraman and Sameer Antani [21]. After using U-Net to identify and extract the lung region in a chest X-ray, the resulting cropped pictures are fed back into the Pre-Trained model to be fine-tuned for improved detection accuracy.

Numerous studies have adopted a mixed-method approach, employing both traditional Data Mining based models like Support Vector Machine (SVM) and Pre-Trained DFE-based models for feature extraction in order to identify Covid-19 infected patients. Sara Hosseinzadeh Kassani et al. [22] presents a DFE based hybrid for computer-assisted decision of Covid-19 pneumonia. Several ML algorithms were trained on deep features extracted by well-known Pre-trained CNNs architectures. Their results suggests that hybrid approach is effective for computer-assisted detection.

#### A. Research Gap

Researchers put forth a lot of time and effort studying how best to detect Covid-19 using deep learning methods. A few research gaps remain, though, as can be seen below.

- Deep learning-based models trained for differentiating pneumonia from Covid-pneumonia fall into the category of Black box models.
- Existing methodologies use complete X-Ray images for detection of disease. X-Ray image contains many body parts other than lung which makes model less reliable.
- Researchers have not considered False Discovery Rate, Negative Predictive Value, False Negative Rate and False Positive Rate while evaluating their proposed model.

This paper resolves these research gaps with the help of U-Net model and Local Interpretable Model-agnostic Explanations methodology.

### III. U-NET

U-Net was originally developed and used in 2015 to process biomedical images; it is an evolution of the conventional convolutional neural network [23]. The network's architecture was improved and expanded from the fully convolutional network to allow it to function with less training images and produce more accurate segmentations. The key concept is to add on more layers to a standard contracting network, with up sampling operators replacing pooling functions. This is because these layers improve the output's resolution. Later convolutional layers can use this data to fine-tune their outputs. U(up sampling) Net's section, which has been modified to include several feature channels, is what allows the network to convey contextual information to finer-grained nodes. This results in a u-shaped layout, with the expanding section mirroring the contracting one.

In this paper, the U-Net architecture is utilized for lung segmentation in patient Chest X-Rays. U-Net model's output for lung segmentation is displayed in Fig. 2. Raw X-Ray image is displayed in Fig. 2(a), while the result of the segmentation process is displayed in Fig. 2(b).

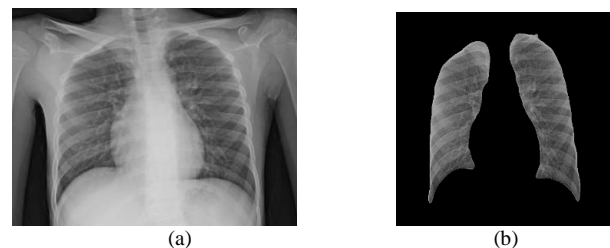


Fig. 2. (a) Original X-Ray Image, (b) Output of U-Net.

From Fig. 2, it is clear that the segmentation process eliminates a great deal of inessential data from the X-Ray image.

### IV. DEEP LEARNING-BASED PRE-TRAINED MODELS

DFE has been widely implemented in the arena of medical imaging in current years. When compared to a CNN that has just been taught, a CNN that has undergone prior training and has undergone appropriate fine-tuning may perform better or on par with the latter. DFE has been vigorously explored for the goal of identifying Covid-19 in chest X-Rays, but there is a dearth of sufficient training data. This section briefly describes the most widely used Pre-Trained models for DFE by researchers.

#### A. MobileNet

It is extremely tiny CNN, making it ideal for integration into small form factors and embedded systems. By using depth-wise distinguishable convolution, a more lightweight architecture is built. Additionally, trade-off hyperparameters are introduced in order to find a fair equilibrium between accuracy and latency [24].

#### B. DenseNet

The term "Densely connected convolutional network" (also known as "DenseNet") refers to the architecture of a network in which every layer is capable of communicating directly with every other layer. Instead of simply adding the feature maps

from each level that came before it, they are now concatenated and added to the most recent layer. As a result of this, DenseNets are able to reuse features while still necessitating a smaller number of parameters than a standard CNN would require to achieve the same level of performance [25].

### C. VGG

The VGG object recognition model is at the bleeding edge of technology and can have as many as 19 layers. VGG, which is a deep CNN, performs better than the baseline on a variety of tasks and datasets, not just ImageNet. The name of the network, VGG, gives away the fact that it is a deep neural network consisting of between 16 and 19 layers [26].

### D. ResNet

Deep residual networks (ResNets), like ResNet-50, are 50-layer CNNs. Using shortcut connections, a residual neural network transforms a plain network. Comparatively, ResNets have fewer filters than VGG Nets [25].

### E. Inception

Inception is a network with a modular architecture made up of recurrent components called Inception units. It gives the inner layers the freedom to determine which filter size is necessary to acquire the necessary knowledge. So the layer adapts to recognize the object no matter what size it is in the image [27].

### F. Xception

By exchanging the traditional Inception components for depthwise Separable Convolutions, Xception augments the inception Style. As with Inception V3, it has large number of parameters [28].

### G. NASNet

ResNet integrated residual learning into the conventional CNN framework. The RU, or residual unit, is made up of a standard layer connected via a skip link. By linking the input of one layer directly to its output, the skip connection facilitates the propagation of signals across the network. That's why it is able to train a super-deep model with the help of RUs [29].

### H. EfficientNet

All three dimensions of depth, width, and resolution are scaled equally using a compound coefficient in this convolutional neural network design and scaling method. The rationale for the compound scaling approach is the intuitive understanding that a larger input image necessitates a more complex network with more layers to broaden the approachable field and additional channels to pick up the finer details in the larger image [30].

## V. PROPOSED ARCHITECTURE

Fig. 3 depicts the proposed architecture. Besides the lungs, Chest X-Rays often show other sections of the body, which might complicate the DFE process and reduce the accuracy of the classifier's predictions. U-Net is applied to the issue of X-Ray image segmentation in order to find a solution. The image of the lungs is kept intact during the segmentation process, but the rest of the picture is completely blacked out as shown in Fig. 2.

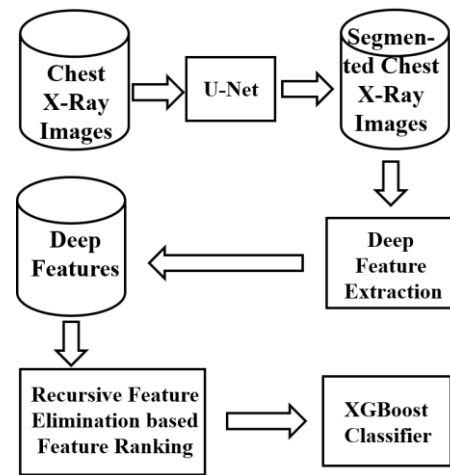


Fig. 3. Proposed Architecture.

Segmented Chest X-Ray images undergo a two-step feature extraction technique. In the first step, a set of segmented X-Ray images is sent to a Pre-Trained model to extract features. As soon as the features have been extracted, they are ranked using Recursive Feature Elimination. At last, it is decided to retain only those features with a ranking of one. This two-step process is described by Algorithm 1 (Deep Feature Extraction).

#### Algorithm 1: Deep Feature Extraction

**Input:** Pre-Trained Model (i.e. M), Set of 'N' Images  $I_1 \dots I_N$ , Required Image Size (i.e.,  $K \times K$ )

**Output:** Matrix Representing Deep Features after RFE, class\_labels

**Procedure:**

$M_R \leftarrow \text{Remove\_Non\_Conv\_Layers}(M)$

$DF = []$

for  $i=1$  to  $N$  do

    Resize image  $I_i$  to size  $(K \times K)$

    features  $\leftarrow M_R(I_i)$

    features  $\leftarrow \text{Flatten}(\text{features})$

$DF \leftarrow DF \cup \text{features}$

Feature-rank  $\leftarrow \text{Feature-Ranking}(DF)$

selected-features =  $DF[\text{feature-rank} == 1]$

In RFE, given an external classifier that assigns weights to features, the primary goal is to determine features by recurrently considering progressively reduced sets of features. The classification model is pre-trained on a limited number of features, and their relative credibility is then established through inspection of a predefined attribute or the execution of a user-defined function. Next, the current set of features is culled to remove the extraneous ones. At last, features are eliminated one by one until only a handful remain.

RFE is a method that relies on a classification model to determine how features should be ranked. Features that aid the classification model in increasing accuracy are ranked higher, whereas features that decrease accuracy are ranked lower. The proposed approach employs the use of REF to guarantee that the output of Pre-Trained models is used to determine the most suitable set of tuning parameters for each variant of the XGBoost classifier model.

The Gradient Boosted decision trees algorithm has been implemented in XGBoost. Here, decision trees are crafted in a logical order. In XGBoost, weights serve a crucial function. All

of the independent variables are given weights, and this information is then used to feed into the decision tree, which makes predictions. Variables that the first decision tree incorrectly predicted are given a higher weight in the second tree's analysis. The ensemble of these separate classifiers/predictors yields a robust and accurate model [31].

Adjustments to the XGBoost Classifier need to be made for a wide range of parameter values so that a precise estimate of inference may be made. In all 29 different parameter values can be tuned for obtaining the best performance of XGBoost model [32]. However, this tuning process requires extremely high processing power in terms of time and computation. Thus, we selected three most widely tuned parameters namely Learning Rate (LR), the number of estimators (classification trees), and the Maximum Depth (MD). The XGBoost Classifier Tuning Algorithm 2 goes into detail about the tuning process. An XGBoost classifier allows the user to modify aspects, including LR, the number of estimators, and MD of each tree.

**Algorithm 2:** Tune XGBoost Classifier

```
Input: Matrix Representing Deep Features after RFE, class_labels  
Output: XGBoost Classifiers' Finest Tuning Parameters  
Procedure:  
Accuracy-Array = []  
for max-depth = 3 to 5 do  
  for learning-rate = 0.1 to 0.5 (step-size 0.2) do  
    for n-estimator = 300 to 800 (step-size 50) do:  
      accuracy-XGB ← XGBoostModelEvaluation  
        (max_depth, learning_rate, n_estimators)  
      Accuracy-Array[n_estimators, learning_rate, max-depth]  
        ← accuracy-XGB  
  Select combinations of Parameters from Accuracy-Array which  
  gives highest accuracy
```

Having a firm grasp on how our machine learning model operates is becoming increasingly crucial. Merely looking at the model's accuracy as a metric of its efficacy is insufficient, though, because it can mislead you. Black-box models, such as deep CNN, are notoriously challenging to interpret. Local Interpretable Model-agnostic Explanations (LIME) aids in comprehending the operation of classifier model. It denotes the visual features that the model employs to classify the images.

Fig. 4(a) shows features derived from an X-Ray image without lung segmentation, while Figure 4(b) shows features extracted from an X-Ray image with the lungs segmented using U-Net. These images demonstrate how features retrieved from segmented images are superior to those extracted from unsegmented images for training a model.

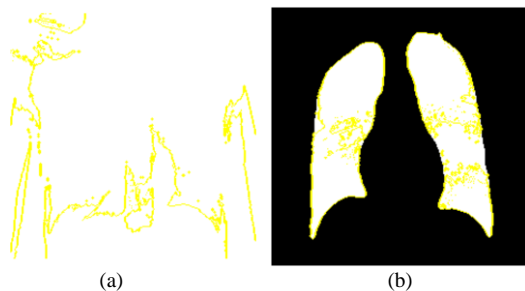


Fig. 4. (a) Feature Extracted from X-Ray Image without Segmentation by Pre-Trained Model, (b) Feature Extracted from X-Ray Image after Segmentation by Pre-Trained Model.

VI. EXPERIMENT DESIGN AND RESULTS

In this section, the data utilized for experimentation along with experimental results is discussed. The April 2021 edition of the COVID-19, Pneumonia and Normal Chest X-ray Dataset is drawn upon for this experimentation [33]. In the first step of data collection process, 613 X-ray images of COVID-19 patients were obtained from the following online resources: GitHub [34] [35], Radiopaedia [36], The Cancer Imaging Archive (TCIA) [37], and the Italian Society of Radiology (SIRM) [38]. Then, rather than the data being individually supplemented, a dataset from Mendeley [39] including 912 photos that had previously been augmented was acquired. Finally, 1525 images of pneumonia cases and 1525 X-ray images of normal cases were retrieved from the Kaggle repository [40] and the NIH dataset [41], respectively. The dataset contains 4575 photos, with 1525 images in each of three distinct groups. The dataset's developer gathered these photographs from a wide range of available web resources.

In order to acquire the utmost precise deep feature, which is decisive to ML, the networks MobileNet, MobileNetLarge, MobileNetSmall, DenseNet121, DenseNet169, DenseNet201, Xception, ResNet50V2, ResNet101V2, ResNet152V2, InceptionV3, InceptionResNetV2, VGGNet16, VGGNet19, NASNet, and EfficientNetB0 through EfficientNetB7 were chosen. Effectiveness of model trained for patient classification based on X-Ray images is evaluated using metrics specified in equations 1 to 5.

$$Precision = \frac{TP}{(TP+FP)} \tag{1}$$

$$Recall = \frac{TP}{(TP+FN)} \tag{2}$$

$$Specificity = \frac{TN}{(TN+FP)} \tag{3}$$

$$Accuracy = \frac{(TP+TN)}{(TP+FP+TN+FN)} \tag{4}$$

$$F1 = 2 * \frac{Precision*Recall}{(Precision+Recall)} \tag{5}$$

For validation of proposed approach, experimentation dataset is divided into two parts namely Training Set and Testing Set. Training Set contains randomly selected 70% images (i.e. 3202) where as remaining 30% images (i.e. 1373) are used to test effectiveness of proposed approach. Evaluation metric for every Pre-Trained model are calculated by taking average of 10 iterations of Training-Testing evaluation in which different set of randomly selected images are used for training and testing the proposed approach.

Fig. 5 depicts the precision of tuned XGBoost models that make use of features retrieved by algorithm 1 as well as a variety of pre-trained models. Evidently, the tweaked XGBoost model that was trained using the features that were retrieved by Algorithm 1 and EfficientNetB1 obtains the best precision of 0.964. Results for precision and other evaluation metrics used in this expropriation for EfficientNetB4 to EfficientNetB7 are less compared to any other models. Thus, we have not presented these results in Fig. 5 to Fig. 9.



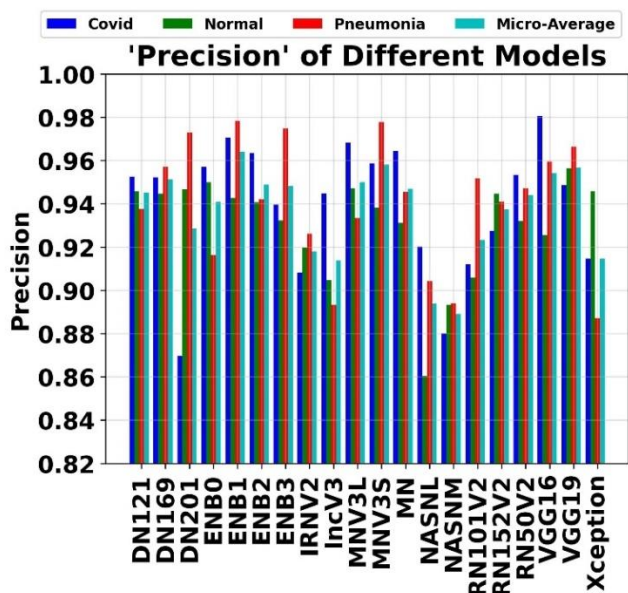


Fig. 5. Precision of Tuned XGBoost Models Trained using Features Extracted by Algorithm 1 (Deep Features Extraction) and Assorted Pre-Trained Models.

Fig. 6 depicts the recall of tuned XGBoost models that make use of features retrieved by algorithm 1 as well as a variety of pre-trained models. Evidently, the tweaked XGBoost model that was trained using the features that were retrieved by Algorithm 1 and EfficientNetB1 obtains the best recall of 0.964.

Fig. 7 depicts the specificity of tuned XGBoost models that make use of features retrieved by algorithm 1 as well as a variety of pre-trained models. Evidently, the tweaked XGBoost model that was trained using the features that were retrieved by Algorithm 1 and EfficientNetB1 obtains the best specificity of 0.982.

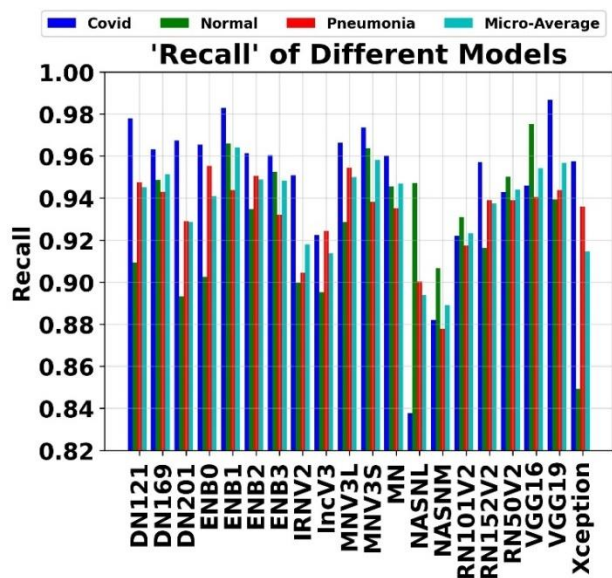


Fig. 6. Recall of Tuned XGBoost Models Trained using Features Extracted by Algorithm 1 (Deep Features Extraction) and Assorted Pre-Trained Models.

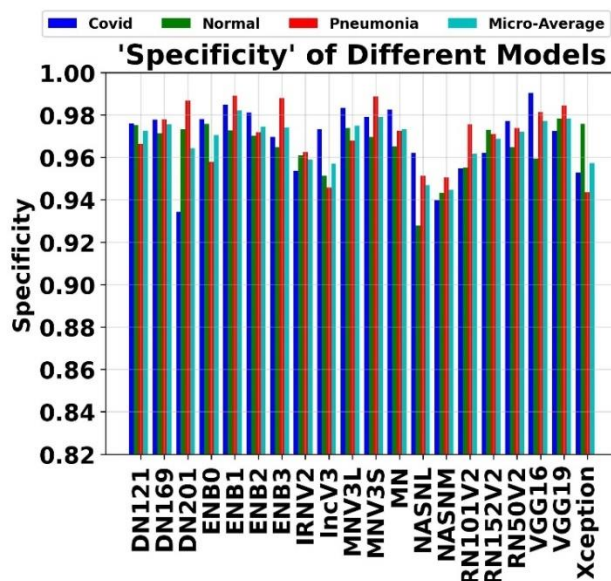


Fig. 7. Specificity of Tuned XGBoost Models Trained using Features Extracted by Algorithm 1 (Deep Features Extraction) and Assorted Pre-Trained Models.

Fig. 8 depicts the accuracy of tuned XGBoost models that make use of features retrieved by algorithm 1 as well as a variety of pre-trained models. Evidently, the tweaked XGBoost model that was trained using the features that were retrieved by Algorithm 1 and EfficientNetB1 obtains the best accuracy of 0.976, which is equivalent to 97.6%.

Fig. 9 depicts the F1-Score of tuned XGBoost models that make use of features retrieved by algorithm 1 as well as a variety of pre-trained models. Evidently, the tweaked XGBoost model that was trained using the features that were retrieved by Algorithm 1 and EfficientNetB1 obtains the best F1-score of 0.964.

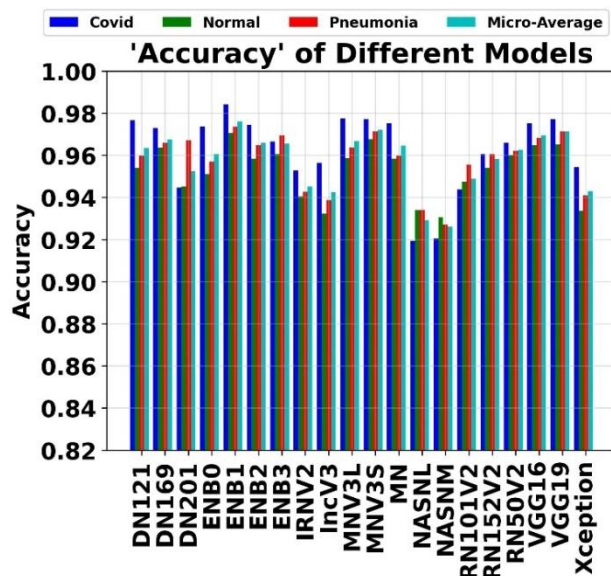


Fig. 8. Accuracy of Tuned XGBoost Models Trained using Features Extracted by Algorithm 1 (Deep Features Extraction) and Assorted Pre-Trained Models.

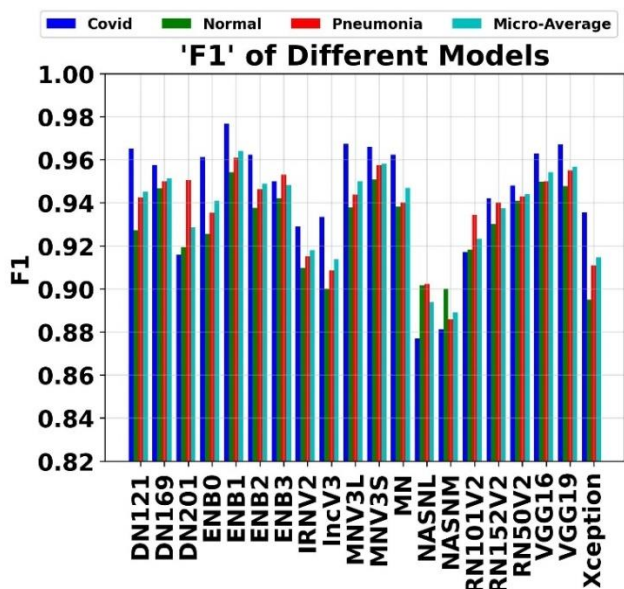


Fig. 9. F1 of Tuned XGBoost Models Trained using Features Extracted by Algorithm 1 (Deep Features Extraction) and Assorted Pre-Trained Models.

Additional set of trained model’s (evolution) metrics are provided in table 1 for fine-tuned XGBoost model that was trained using the features that were retrieved by Algorithm 1 and EfficientNetB1. Equations for these metrics are given below in equation 6 to equation 9. These metrics are provided for individual categories i.e. Covid-19, normal and Pneumonia and well as overall testing dataset.

$$\text{False Discovery Rate} = \frac{FP}{(FP+TP)} \tag{6}$$

$$\text{Negative predictive value} = \frac{TN}{(TN+FN)} \tag{7}$$

$$\text{False Negative Rate} = \frac{FN}{(FN+TP)} \tag{8}$$

$$\text{False Positive Rate} = \frac{FP}{(FP+TN)} \tag{9}$$

TABLE I. EVALUATION METRICS FOR FINE-TUNED XGBOOST MODEL THAT WAS TRAINED USING THE FEATURES THAT WERE RETRIEVED BY ALGORITHM 1 AND EFFICIENTNETB1

Metric	Covid-19	Normal	Pneumonia	Micro-Average
Accuracy	98.4%	97.1%	97.3%	97.6%
F1	0.977	0.954	0.961	0.964
Precision	0.971	0.943	0.979	0.964
Recall	0.983	0.966	0.944	0.964
Specificity	0.985	0.973	0.989	0.982
False Discovery Rate	0.029	0.057	0.021	0.036
Negative Predictive Value	0.991	0.984	0.971	0.982
False Negative Rate	0.017	0.034	0.056	0.036
False Positive Rate	0.015	0.027	0.011	0.018

TABLE II. COMPARISON OF PROPOSED METHODOLOGY WITH RELATED WORK

Reference	Method	Dataset	Accuracy
[1]	22-Layer CNN	Merged Publicly Available Datasets	94.2%
[9]	VGG16 + CLAHE	Merged Publicly Available Datasets	88.8%
[10]	Xception + ResNet50V2	Merged Publicly Available Datasets	91.4%
[11]	DenseNet-121	Merged Publicly Available Datasets	81.04%
[12]	ResNet50V2 + VGG-16 + InceptionV3	Merged Publicly Available Datasets	95.49%
[13]	VGG16	Publicly available Dataset	91.69%
[14]	COVID-GATNet	Merged Publicly Available Datasets	94.1%
[15]	VGG16	Publicly available Dataset	91.0%
[16]	DenseNet169	Publicly available Dataset	95.72
[17]	AlexNet	Merged Publicly Available Datasets	94.0%
[18]	Xception	Merged Publicly Available Datasets	89.6%
[20]	EfficientNetB3	Merged Publicly Available Datasets	93.9%
Proposed	U-NET + EfficientNetB1 + XGBoost	Merged Publicly Available Datasets	97.6%

Table II details how the proposed architecture and experimental results compare to the state-of-the-art. Because there is no universally accepted reference dataset, it is clear that all researchers are making use of data that is already available publicly.

## VII. DISCUSSION

Significant research contributions of this paper are as follows:

1) Proposed use of U-Net model to segment lung part from X-Ray image and then segmented image to extract deep features.

2) Proposed model is trained on reduced features set using XGBoost classifier model, thus it can be easily deployed on low-end cost-efficient devices like Raspberry-pi easily.

3) Use of Explainable AI: Local Interpretable Model-agnostic Explanations (LIME) methodology is used to explain the decision-making process of Deep Learning Model (Explained in section V).

4) Performance of 20 different Pre-Trained models is evaluated against 9 different metrics specified in Table I.

Limitations of the Research work carried out are as follows:

1) As discussed in Section V, the XGBoost model has 29 parameters that can be adjusted to optimize performance. This study, however, relied on only three variables. This allows for

additional investigation into how the remaining XGBoost classifier parameters affect performance.

2) Using the U-Net model for segmentation does not result in any image enhancement of the lung segment area. Pre-Trained models for deep feature extraction can be used after various image enhancement techniques on lung segmented images have been performed.

### VIII. CONCLUSION

In this study, a Deep Feature Extraction-based method for detecting individuals infected with Covid-19 and pneumonia by analyzing Chest X-Rays is provided. U-Net is used to split the lung region of the chest from the image so that the Pre-Trained model may concentrate on the relevant area of the image and extract more meaningful features. In this article, following the segmentation of the lung part, a two-stage technique to the extraction of deep features from X-Ray images is provided. After that, the XGBoost classifiers are trained to exploit these features in order to differentiate between patients infected with Covid-19 and healthy individuals and patients infected with pneumonia. When the performance of 20 different Pre-Trained models is analyzed, it is discovered that the maximum detection accuracy, precision, recall, specificity, and F1-score are achieved when EfficientNetB1 is used to extract deep features. The respective values for these metrics are 97.6%, 0.964, 0.964, and 0.982. These findings lend credence to the efficiency of the strategy that was proposed.

### REFERENCES

- [1] E. Hussain, M. Hasan, M. A. Rahman, I. Lee, T. Tamanna, and M. Z. Parvez, "CoroDet: A deep learning based classification for COVID-19 detection using chest X-ray images," *Chaos, Solitons and Fractals*, vol. 142, 2021, doi: 10.1016/j.chaos.2020.110495.
- [2] "https://www.worldometers.info/coronavirus/worldwide-graphs/".
- [3] A. M. Ismael and A. Şengür, "Deep learning approaches for COVID-19 detection based on chest X-ray images," *Expert Syst. Appl.*, vol. 164, 2021, doi: 10.1016/j.eswa.2020.114054.
- [4] A. Maier, C. Syben, T. Lasser, and C. Riess, "A gentle introduction to deep learning in medical image processing," *Zeitschrift für Medizinische Physik*, vol. 29, no. 2, pp. 86–101, 2019, doi: 10.1016/j.zemedi.2018.12.003.
- [5] K. Weiss, T. M. Khoshgoftaar, and D. D. Wang, "A survey of transfer learning," *J. Big Data*, vol. 3, no. 1, 2016, doi: 10.1186/s40537-016-0043-6.
- [6] E. Deniz, A. Şengür, Z. Kadiroğlu, Y. Guo, V. Bajaj, and Ü. Budak, "Transfer learning based histopathologic image classification for breast cancer detection," *Heal. Inf. Sci. Syst.*, vol. 6, no. 1, 2018, doi: 10.1007/s13755-018-0057-x.
- [7] Z. Li, F. Liu, W. Yang, S. Peng, and J. Zhou, "A Survey of Convolutional Neural Networks: Analysis, Applications, and Prospects," *IEEE Trans. Neural Networks Learn. Syst.*, pp. 1–21, 2021, doi: 10.1109/tnnls.2021.3084827.
- [8] C. Ieracitano et al., "A fuzzy-enhanced deep learning approach for early detection of Covid-19 pneumonia from portable chest X-ray images," *Neurocomputing*, vol. 481, pp. 202–215, Apr. 2022, doi: 10.1016/j.neucom.2022.01.055.
- [9] F. Saiz and I. Barandiaran, "COVID-19 Detection in Chest X-ray Images using a Deep Learning Approach," *Int. J. Interact. Multimed. Artif. Intell.*, vol. 6, no. 2, p. 4, 2020, doi: 10.9781/ijimai.2020.04.003.
- [10] M. Rahimzadeh and A. Attar, "A modified deep convolutional neural network for detecting COVID-19 and pneumonia from chest X-ray images based on the concatenation of Xception and ResNet50V2," *Informatics Med. Unlocked*, vol. 19, 2020, doi: 10.1016/j.imu.2020.100360.
- [11] A. Haghaniifar, M. M. Majdabadi, Y. Choi, S. Deivalakshmi, and S. Ko, "COVID-CXNet: Detecting COVID-19 in frontal chest X-ray images using deep learning," *Multimed. Tools Appl.*, vol. 81, no. 21, pp. 30615–30645, Sep. 2022, doi: 10.1007/s11042-022-12156-z.
- [12] M. Shorfuzzaman, M. Masud, H. Alhumyani, D. Anand, and A. Singh, "Artificial Neural Network-Based Deep Learning Model for COVID-19 Patient Detection Using X-Ray Chest Images," *J. Healthc. Eng.*, vol. 2021, 2021, doi: 10.1155/2021/5513679.
- [13] M. D. K. Hasan et al., "Deep Learning Approaches for Detecting Pneumonia in COVID-19 Patients by Analyzing Chest X-Ray Images," *Math. Probl. Eng.*, vol. 2021, 2021, doi: 10.1155/2021/9929274.
- [14] J. Li, D. Zhang, Q. Liu, R. Bu, and Q. Wei, "COVID-GATNet: A Deep Learning Framework for Screening of COVID-19 from Chest X-Ray Images," in *2020 IEEE 6th International Conference on Computer and Communications, ICC3 2020*, 2020, pp. 1897–1902, doi: 10.1109/ICC351575.2020.9345005.
- [15] D. Dansana et al., "Early diagnosis of COVID-19-affected patients based on X-ray and computed tomography images using deep learning algorithm," *Soft Comput.*, 2020, doi: 10.1007/s00500-020-05275-y.
- [16] K. Hammoudi et al., "Deep Learning on Chest X-ray Images to Detect and Evaluate Pneumonia Cases at the Era of COVID-19," *J. Med. Syst.*, vol. 45, no. 7, 2021, doi: 10.1007/s10916-021-01745-4.
- [17] A. U. Ibrahim, M. Ozsoz, S. Serte, F. Al-Turjman, and P. S. Yakoi, "Pneumonia Classification Using Deep Learning from Chest X-ray Images During COVID-19," *Cognit. Comput.*, 2021, doi: 10.1007/s12559-020-09787-5.
- [18] A. I. Khan, J. L. Shah, and M. M. Bhat, "CoroNet: A deep neural network for detection and diagnosis of COVID-19 from chest x-ray images," *Comput. Methods Programs Biomed.*, vol. 196, 2020, doi: 10.1016/j.cmpb.2020.105581.
- [19] H. Panwar, P. K. Gupta, M. K. Siddiqui, R. Morales-Menendez, P. Bhardwaj, and V. Singh, "A deep learning and grad-CAM based color visualization approach for fast detection of COVID-19 cases using chest X-ray and CT-Scan images," *Chaos, Solitons and Fractals*, vol. 140, 2020, doi: 10.1016/j.chaos.2020.110190.
- [20] E. Luz et al., "Towards an effective and efficient deep learning model for COVID-19 patterns detection in X-ray images," *Res. Biomed. Eng.*, vol. 38, no. 1, pp. 149–162, Mar. 2022, doi: 10.1007/s42600-021-00151-6.
- [21] S. Rajaraman and S. Antani, "Weakly labeled data augmentation for deep learning: A study on COVID-19 detection in chest X-rays," *Diagnostics*, vol. 10, no. 6, 2020, doi: 10.3390/diagnostics10060358.
- [22] S. H. Kassania, P. H. Kassanib, M. J. Wesolowskic, K. A. Schneidera, and R. Detersa, "Automatic Detection of Coronavirus Disease (COVID-19) in X-ray and CT Images: A Machine Learning Based Approach," *Biocybern. Biomed. Eng.*, vol. 41, no. 3, pp. 867–879, 2021, doi: 10.1016/j.bbe.2021.05.013.
- [23] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2015, vol. 9351, pp. 234–241, doi: 10.1007/978-3-319-24574-4\_28.
- [24] D. Sinha and M. El-Sharkawy, "Thin MobileNet: An Enhanced MobileNet Architecture," in *2019 IEEE 10th Annual Ubiquitous Computing, Electronics and Mobile Communication Conference, UEMCON 2019*, 2019, pp. 0280–0285, doi: 10.1109/UEMCON47517.2019.8993089.
- [25] C. Zhang et al., "ResNet or DenseNet? Introducing dense shortcuts to ResNet," in *Proceedings - 2021 IEEE Winter Conference on Applications of Computer Vision, WACV 2021*, 2021, pp. 3549–3558, doi: 10.1109/WACV48630.2021.00359.
- [26] S. Pasban, S. Mohamadzadeh, J. Zeraatkar-Moghaddam, and A. K. Shafiei, "Infant brain segmentation based on a combination of vgg-16 and u-net deep networks," *IET Image Process.*, vol. 14, no. 17, pp. 4756–4765, 2020, doi: 10.1049/iet-ipr.2020.0469.
- [27] Z. Ying et al., "Tai-sarnet: Deep transferred atrous-inception cnn for small samples sar atr," *Sensors (Switzerland)*, vol. 20, no. 6, 2020, doi: 10.3390/s20061724.

- [28] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, 2017, vol. 2017-Janua, pp. 1800–1807, doi: 10.1109/CVPR.2017.195.
- [29] F. Martínez, F. Martínez, and E. Jacinto, "Performance evaluation of the NASnet convolutional network in the automatic identification of COVID-19," Int. J. Adv. Sci. Eng. Inf. Technol., vol. 10, no. 2, pp. 662–667, 2020, doi: 10.18517/ijaseit.10.2.11446.
- [30] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in 36th International Conference on Machine Learning, ICML 2019, 2019, vol. 2019-June, pp. 10691–10700.
- [31] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, vol. 13–17–Augu, pp. 785–794, doi: 10.1145/2939672.2939785.
- [32] "No Title." <https://xgboost.readthedocs.io/en/stable/parameter.html>.
- [33] Z. Asraf, Amanullah; Islam, "COVID19, Pneumonia and Normal Chest X-ray PA Dataset," Mendeley Data, V1, 2021, doi: 10.17632/jctsfj2sf.1.
- [34] <http://arxiv.org/abs/2003.11597>.
- [35] <https://github.com/agchung>.
- [36] <https://radiopaedia.org/>.
- [37] <https://www.cancerimagingarchive.net/>.
- [38] <https://www.sirm.org/en/category/articles/covid-19-database/>.
- [39] <https://data.mendeley.com/datasets/2fxz4px6d8/4>.
- [40] <https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia>.
- [41] <https://www.kaggle.com/nih-chest-xrays/data>.

# Analysis of Privacy and Security Challenges in e-Health Clouds

Reem Alanazi

Faculty of Engineering and Information Technology  
University of Technology Sydney, Australia

**Abstract**—Electronic Health Records (EHR) techniques are being used at an increasingly faster rate to store the patient data making it easier to retrieve, share and utilize it efficiently. This data can be used for research purposes, clinical trials and for studying epidemiology to come up with strategies for epidemic control. With a huge global inflation, the increasing costs of healthcare and the shortage of medicine, it becomes convenient for the healthcare organizations to migrate from the traditional healthcare system to a more sophisticated, cost effective and efficient cloud-based e-Health model. To optimize the full potential of an ICT-based e-Health system, it is imperative for the existing healthcare systems to be implemented in a full-fledged cloud environment. However, with numerous benefits of technology, it might pose some privacy and security threats as well. Therefore, the security and access control of such information is of vital significance. Nonetheless, with the increasing interest of healthcare organizations to migrate from the conventional healthcare systems to the modern cloud-based e-Health systems, not much care is being taken to address security and privacy issues effectively towards the protection of sensitive data.

**Keywords**—HER; e-health; security; privacy; cloud

## I. INTRODUCTION

With the advancements in Information and Technology, services all across the globe have improved, in all fields in general, and around healthcare in particular, all due to several provisions like flexible processes and low-cost treatments. When healthcare is integrated with the internet it is termed as e-Health [1] [2]. This e-Health is the biggest innovation of technology and its implementation in a cloud-based environment is necessary to maximize the benefits. Despite all the benefits that e-health has been providing, it nevertheless still gets affected by certain challenges of privacy and security [1]. Since its conception, cloud computing has garnered enough popularity around the health sector. The cloud-based e-health makes the sharing of this health data with its stakeholders easier [3]. Although the internet by making use of technologies like cloud computing has made the concept of centralized healthcare possible, it has however brought in with it certain bugs and loopholes in the system. There are certain issues concerning security and privacy that remain unaddressed and unresolved even today [3].

The services that the e-health cloud offers are preventive care, keeping an account of patient satisfaction, a continuous vigil and AI supported detection. All these services are prone to privacy breach, and it needs to be taken care while rules regarding privacy measures are implemented. With the growth

of use of technology among people, the awareness regarding privacy of their medical data has grown as well. Patients fear leak of their private medical history that they might be embarrassed of, to social media. It's important to maintain the trust of patients in health services providers, thus the e-health cloud needs to be highly secured [2] [9]. The centralization and digitization of the data in the health sector has made the sharing of medical data easy. However, this sharing might lead to data attacks and cause loss of confidential data. To tackle this, several government bodies have taken initiatives to ensure better security and proper privacy of data. Instances from the US healthcare industry show the progress that's been made in securing the e-health systems. The Health Insurance Portability and Accountability Act (HIPAA), which had guidelines set for security and privacy requirements of US healthcare, ensured proper utilization of e-health [1]. A secure e-health system ensures the system has these attributes; Authenticity, Integrity, Availability, Access control, Anonymity [4].

## II. LITERATURE REVIEW

Several articles have been written and read about e-Health and its merits and demerits. The research has been done about what challenges are being faced in the pursuit of better security and maintaining privacy of data, and how these drawbacks can be addressed. Security models developed in relation with healthcare applications, targeted the information loss and ways to combat it. A security model, Role Based Access Control (RBAC) was deployed to find solutions for the already identified security challenges in electronic health [31]. An extended version of RBAC known as u-healthcare, was designed to carry out four vital functions: what meal should the patient take, exchange of information related to health, management of the same health information on smart devices that the records are accessed through. These studies however concluded not much could be resolved about these security issues. The model had several drawbacks, and it was not suitable for disturbed environment. The solution offered had limited application. It was not scalable to any number of users.

Another model developed by [32] demonstrated the working of a comparatively lightweight framework that was based on Transport Layer Security/Secure Sockets Layer (TLS/SSL) to secure the data shared between the server and client. This framework was called Secure Health and it had many security features like proper authentication, and authorization for the transmitted and stored data. It protected the system from unidentified and unauthorized access minimizing alien access to confidential health data. It also

helped the administrator in identifying wrongly stored information [33]. Despite all the benefits that the system offered, some challenges remained unresolved, the main one being its platform dependency and no or less scalability. In cloud computing, scalability is the most sought feature. The systems based on cloud should have provision for an expansion in future. In the need to minimize the cost of maintaining the health data and for keeping it available but secure, another security mechanism with a different concept of hierarchy was given by Barua, et al. This model adopted the theory of Attribute Based Encryption (ABE). It had provision for access control, which was framed at the central level. However, even this approach failed at addressing a large number of requests from the client side because of the centralized manner in which the health data was stored in [34]. In case of many users accessing the system at once, the requests needed to be sorted out depending on their priorities. To overcome the challenges caused due to the centralization of data, other attributes like collaborative and distributive nature of the e-health systems were taken in consideration by Guo et al. [35]. They brought about a change in the way requests were being addressed and servers being accessed. They suggested authorization from both the patients and the doctors, and not from the centralized server itself. Clients were given access only on the basis of priorities concealing their identities and characteristics. Hamid et al. worked on the confidentiality of patient's multimedia data in the cloud. They proposed a bilinear pairing cryptography based triparty one-round authenticated key agreement protocol. This protocol helps in secure communication by generating a session key. Also, a decoy technique with a fog computing facility has been implemented so that the private healthcare data can be accessed and secured securely. This approach induces computational expenses in communication for strong security [10-15]. Marwan et al. proposed a new method to enhance the reliability of cloud storage and used Shamir's Secret Share Scheme (SSS) and multi cloud concept. This was done to meet security requirements, avoid data loss, and prevent unauthorized access and privacy disclosure. To prevent medical records from getting leaked and revealed, this technique allows division of the data into small shares. Also, data is spread among various cloud storage systems. Medical Data is encrypted using SSS technique and split into shares to ensure confidentiality and privacy. This article has not discussed any aspect of the optimal number of shares aroused to tradeoff between efficiency and security [16, 17]. Galletta et al. have proposed a system developed at Istituto di Ricovero e Cura a Carattere Scientifico (IRCCS) claiming to address security and privacy of patient's data [18]. This system works on two software parts: the splitter and anonymizer. The anonymizer collects anonymized clinical data and the splitter obscures and stores health data in multiple cloud storage providers. This data can be accessed by only authorized clinical operators. Moreover, the performance of the system has been assessed by magnetic resonance imaging (MRI) data. Alexander et al. proposed anonymization techniques and privacy-aware systems in order to publish data on the cloud. This system used Advanced Encryption Standard (AES) and k-anonymity [19-21]. Smithamol et al. proposed a novel architecture to address data confidentiality. This model

constructed the group-based access structure and Ciphertext-Policy Attribute-Based Encryption (CP-ABE) by making use of the partially ordered set, thereby providing medical records access control. This approach reduced overall encryption time and computational overhead [22] [23]. Ibrahim et al. provided a solution to access securely the privacy-sensitive EHR data through: 1) a cryptographic role-based technique for the distribution of session keys with the help of Kerberos protocol, 2) location and biometrics-based process for user authorization, and 3) a wavelet-based steganographic technique for embedding EHR data securely. This approach showed resilience to man in the middle attack and replay attacks. However, it did not show resilience to other security threats and did not analyze its scalability [24]. Shah and Prasad deployed a novel structure with cloud-based privacy-aware role-based access control (CPRBAC) model. This model presented a list of different methods of encryption and various privacy, and security challenges were addressed. Its goal was to minimize computational complexity. However, no qualitative analysis was carried out to check the efficiency of the approach [25]. Supriya and Padaki surveyed various lapses in health care security particularly concerned with non-repudiation, CIA model. They studied and discussed some already proven operational strategies and methodologies related to risk management. This way they were able to perceive what the health industry must follow to reduce security and privacy threats [26]. Lohr et al. tried to establish privacy domains in e-health infrastructures by presenting a security architecture based on Trusted Virtual Domains (TVDs). This architecture, however, did not address other research challenges like anonymity, non-repudiation, incapacity of the patient to authenticate [27, 28]. Kumar et al. proposed a model based on encryption technique called Attribute Based Encryption (ABE). All the users have been divided into two domains: personal and public. This is done to control key management complexity. In the personal domain, a user can encrypt data that is allocated to him whereas in the public domain, a user can adopt and utilize multi-authority ABE. Scalability and flexibility are the two challenges associated with this approach as integrating ABE into the EHR system gives rise to serious and key management challenges [29]. Zhu et al. proposed a model that utilized re-encryption and Attribute Based Encryption (ABE) with proxy encryption which is enabled by Rivest Shamir and Adleman (RSA). The whole purpose of using proxy encryption was to induce a separation mechanism to validate the patient's data. Write privilege keys were given to professionals whereas read privilege keys were given to patients. Using this model, computation overhead was minimized. The healthcare worker can be easily stopped from getting the read keys and this does not need to be approved from both ends. Moreover, this framework can be accessed by only a few users [30].

### III. WHAT IS E-HEALTH

The concept of conventional healthcare has been narrowed down to virtual healthcare. The integration of technology with healthcare has emerged as a new concept termed as e-health. In an era like today time management is the need of the hour. Going to doctors for consultations and waiting in queues for hours is an obsolete scene now. With the internet and



technology being used in all walks of life, people prefer a virtual way of seeking medical treatment. The process of having consultations online, using desktops and laptops is e-Health [5]. This cloud-based e-health system can be created specifically according to the needs of the patient. Also, healthcare accessed via mobile phones is m-health which is mostly for self-management of chronic diseases. The e-Health systems have been of great help in dealing with patients who had prolonged diseases and needed to be kept constant vigil. These e-Health systems keep a track of health-related information and provide help with symptom checking, finding a suitable doctor, managing financial records, self-monitoring, and filling prescriptions. This information is available and accessed by a large population [5]. E-health has now become a vital part of the healthcare system due to its efficient services and accurate results, and error free unlike the traditional healthcare systems. In a traditional way of treatment, a patient could get a particular dose of medicine twice due to manual handling of records. This is not the case with health where electronic medical records are maintained, which store all the information about the patient's treatment and medicines given, thus avoiding any errors with the medication of the patient [6]. A country's success of e-Health is derived from many factors like what type of management and infrastructure is being used, how much is the user involvement, if the system is scalable to adapt to as many users as possible or not. The medical data in the cloud is of importance to its healthcare professionals, patients, entrepreneurs, and businesses which deal with health insurances or such policies. E-health strategies like norms, laws and regulations must be created to implement cloud technology in healthcare effectively. It is not just another

progress in the field of technology, it is a way of thinking, state of mind, to increase the reach of healthcare, improving local, regional and global healthcare by making use of ICT (Information and Communication Technology).

As the name suggests the e-Health uses electronics which can be mobile phones or computers along with cloud technology for storage purposes. The eHealth is mainly of two types: Personal Health records (P-HR) and Electronic Health Records (E-HR). The P\_HR is used by patients to update their health records themselves and for seeking consultations online by using mobile phones. Whereas E-HR is of use to healthcare professionals. This E-HR is very beneficial in providing the right treatment to the patient and for secure sharing medical records or patient medical history, medicine details and prescriptions. E-health provides an effective and real time treatment for the patient.

#### IV. CLOUD BASED E-HEALTH MODELS

1) *Private cloud*: Model given in Fig. 1 is the most secured model. It is usually operated and managed by the same organization that uses it, making it highly secured and hence security issues are limited. It is located on premises, over the intranet and behind the firewall. In this, the public internet is completely restricted. Only recognized personnel from the healthcare institution are able to access the electronic medical record (EMR). Only these personnel are considered to be trustworthy and reliable. A good example is VMware [1] [4].

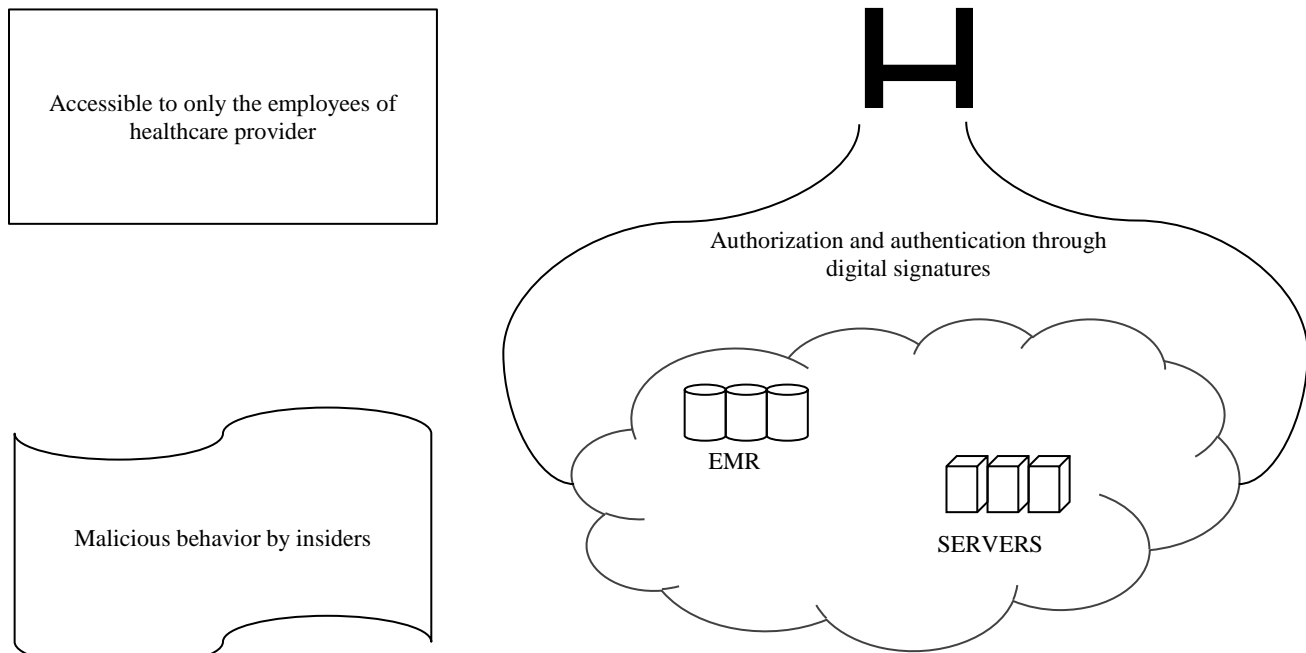


Fig. 1. Private Cloud.

2) *Public cloud*: This model given in Fig. 2 is totally in control of the third-party provider as the services of the cloud system are provided by them only. They are known as cloud service providers (CSPs). It is located off premises, over the internet controlled by CSPs. In this, EHRs are held between various organizations and these EHRs are highly vulnerable to malicious attacks. It has many security challenges associated with it and to avoid them, efficient cryptographic mechanisms and fine-grained access control frameworks need to be applied. It is considered less secure than the private cloud. Some good examples are Dropbox, Amazon EC2, Microsoft Azure [1] [4].

3) *Hybrid cloud*: Model given in Fig. 3 is a combination of public and private cloud servers where both works

individually but unitedly. The deployment of this model combines the benefits of both the models and multiple cloud services can be used. This model is highly advantageous to e-health and plays an important role in integration, composition and organizational impact and housing big medical data. Health care providers that have restricted and limited resources can easily deploy this model which makes use of third-party services. There is a huge importance of hybrid cloud in health care but at the same time there is a need of effectively implementing hybrid cloud in e-health as it has trust and confidentiality issues because of the public part. An important example is Rackspace [1] [7].

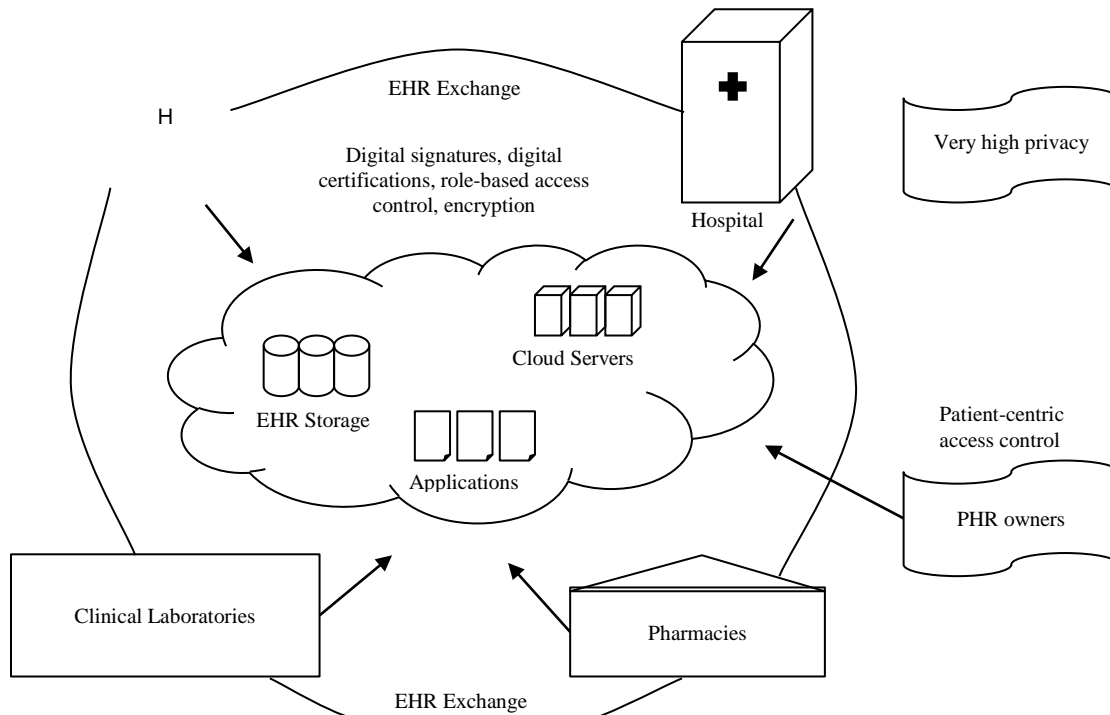


Fig. 2. Public Cloud.

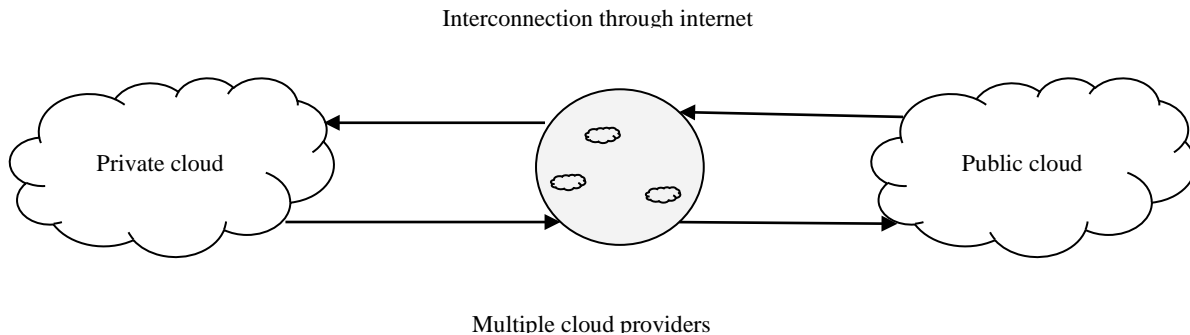


Fig. 3. Hybrid Cloud.

## V. PRIVACY AND SECURITY REQUIREMENTS

It is a known fact that transition of conventional health care systems to e-health care has brought with it a lot of benefits and has made health care systems easier and affordable but there are a lot of challenges associated with it like confidentiality, privacy and security of patients' records. Cloud computing is highly acceptable in digital technology which is being used in the healthcare industry extensively [8]. In order to increase the confidence of patients and organizations, the cloud service providers and other government organizations have formulated a range of security measures and guidelines. Cloud servers have been broadly classified into three categories: trusted, semi-trusted, and untrusted. A trusted server can be defined as the one that can be trusted completely. It doesn't lead to any information leak and threats to the health data. Semi trusted servers are those that are considered to be honest but curious servers. They conspire with malicious users to acquire health data. Untrusted servers cannot be trusted and are highly vulnerable to attacks.

In e-health system there is a need of security and privacy in the following ways:

- 1) *Data integrity*: It is a mechanism which ensures health data is not being altered by an unauthorized entity.
- 2) *Data confidentiality*: It is a mechanism which ensures that the sensitive health information does not reach unauthorized users. Data confidentiality is achieved by data encryption.
- 3) *Authenticity*: It is a mechanism that ensures sensitive health data is accessed by only authorized and authentic authority.
- 4) *Accountability*: It is a mechanism to justify the actions and decisions of organizations and individuals.
- 5) *Audit*: It keeps the track of any kind of activity on the health data and is continuously monitored and protected. It also ensures data privacy and security.
- 6) *Non-repudiation*: It refers to the sender and receiver's non denial of authenticity. It means, neither patients nor doctors can repudiate health data after its theft.
- 7) *Anonymity*: It is a mechanism which keeps the identity of the users anonymous so that the cloud servers are unable to access the stored health data.

## VI. CONCLUSION

The security and privacy of patient data in e-Health systems is quite a demanding task. Researcher is being carried out internationally to provide and protect the Electronic Health Records (EHR) data. To mitigate the hurdles of security and privacy challenges, it is essential for health organizations to migrate from the traditional e-Health systems to the advanced cloud-based e-Health systems. While migration towards the cloud-based infrastructure protects the patient data in EHRs to a larger extent, it, however, does not guarantee the full-proof security and protection against data theft and other type of

threats and vulnerabilities. An in-depth analysis has been done in this perspective. Firstly, an attempt has been made to understand the requirement of shifting from traditional e-Health systems to cloud-based infrastructure. A survey has been carried out to highlight the privacy and security considerations prevalent in the cloud-based e-Health systems [8]. More than 30 research papers were analyzed to highlight the underlying security and privacy issues in various cloud-based e-Health systems across the globe. Some security techniques with their pros and cons have also been presented. Finally, some recommendations have been made for enhanced privacy and protection in the cloud-based e-Healthcare infrastructure. For the future research, we are looking forward to exploring the state-of-the-art techniques for preserving privacy especially in terms of EHR data in the cloud infrastructure scenarios.

## VII. FUTURE DIRECTIONS

There has been a tremendous development and progress in security and privacy in e-health clouds. Still, there is a need to enhance and enforce certain security and privacy measures in the e-health system. This can be done by enhancing and maintaining efficiency of all the initiatives regarding security and privacy.

Some of the strategies for security purpose in e-health are mentioned below:

- 1) Auditing is the first thing in assuring security and privacy in e-health. This approach greatly helps in locating and identifying any kind of wrongdoing in the e-health system. Hence, auditing can be regarded as a new research direction for e-health.
- 2) Encryption schemes can also help in achieving security and privacy. Encrypting information that has the data regarding e-health users will in no way lead to insecurity of data. This is an excellent procedure as the other parts of information remain unencrypted.
- 3) RBAC is a model that has been put into use in order to address the issues of security and privacy in e-health. Also, Attribute Based Access Control (ABAC) model is widely used to ensure remarkable scalability and flexibility for authorizations and authentications.
- 4) Attribute Based Encryption (ABE) is an exceptional way to ensure privacy in e-health. But the performance is affected while decrypting data because of bi-linear pairing operations. These bi-linear operations require a solution to strengthen the efficiency of ABE.
- 5) General enforcement needs to be adopted to ensure privacy. Privacy must be incorporated in e-health that benefits all the parties. Also, research must be done to know about the security and privacy violations.
- 6) Most of the solutions adopted RBAC, MAC, and DAC. These models can ensure better results of security and privacy when applied collectively or hybridized into a single model.

REFERENCES

- [1] N. A. Azeez, & C. V. der Vyver, (2018). Security and privacy issues in e-health cloud-based system: A comprehensive content analysis. *Egyptian Informatics Journal*. doi:10.1016/j.eij.2018.12.001.
- [2] C. B. Pheng, K.H. Yeh, & H. Xiong, (2020). Security and Privacy in IoT-Cloud-Based e-Health Systems—A Comprehensive Review. *Symmetry*, 12(7), 1191. doi:10.3390/sym12071191 XX.
- [3] S. Aqeel et.al(2021).A Review of the State of the Art in Privacy and Security in the eHealth Cloud, Doi:/10.1109/ACCESS.2021.3098708,IEEE Access.
- [4] Y. Al-Issa, M. A. Ottom, & A. Tamrawi, (2019). eHealth Cloud Security Challenges: A Survey. *Journal of Healthcare Engineering*, 2019, 1–15. doi:10.1155/2019/7516035.
- [5] L. Leung, & C. Chen, (2019). E-health/m-health adoption and lifestyle improvements: Exploring the roles of technology readiness, the expectation-confirmation model, and health-related information activities. *Telecommunications Policy*. doi:10.1016/j.telpol.2019.01.005.
- [6] M. H. da Fonseca, (2021). E-Health Practices and Technologies: A Systematic Review from 2014 to 2019 ,doi:. <https://doi.org/10.3390/healthcare9091192>.
- [7] H. Kunwal, B. H. Malik, A. Saeed, H. Mushtaq, H. B. Cheema, F. Mehmood, (2017). “Medicloud: Hybrid Cloud Computing Framework to Optimize E-Health Activities”, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 8, No. 9, 2017.
- [8] S. Chenthara, K. Ahmed, H. Wang, F. Whittaker (2019), “Security and Privacy-Preserving Challenges of e-Health Solutions in Cloud Computing”, IEEE, 2019.
- [9] D. K. Yadav, S. Behera, (2020). “A Survey on Secure Cloud-Based E-Health Systems”,.; doi: 10.4108/eai.13-7-2018.163308.
- [10] H. A. Al Hamid, S. M. M. Rahman, M. S. Hossain, A. Almogren, and A. Alamri, “A security model for preserving the privacy of medical big data in a healthcare cloud using a fog computing facility with pairing-based cryptography,” *IEEE Access*, vol. 5, 2017.
- [11] N. Koblitz and A. Menezes, “Pairing-based cryptography at high security levels,” *Cryptography and Coding*, vol. 3796, pp. 13–36, 2005.
- [12] J. Voris, J. Jermyn, A. D. Keromytis, and S. J. Stolfo, “Bait and snitch: defending computer systems with decoys,” in *Proceedings of the Cyber Infrastructure Protection Conference, Strategic Studies Institute, Arlington, VA, USA, September 2013*.
- [13] S. P. Karekar and S. M. Vaidya, “Perspective of decoy technique using mobile fog computing with effect to wireless environment,” *International Journal of Scientific Engineering and Technology Research*, vol. 4, no. 14, pp. 2620–2626, 2015.
- [14] J. Shropshire, “Extending the cloud with fog: security challenges & opportunities,” in *Proceedings of the Americas Conference on Information Systems, AMCIS 2014, Savannah, GA, USA, August 2014*.
- [15] K. Manreet and B. Monika, “Fog computing providing data security: a review,” *International Journal of Computer Science and Software Engineering*, vol. 4, no. 6, pp. 832–834, 2014.
- [16] M. Marwan, A. Kartit, and H. Ouahmane, “Protecting medical data in cloud storage using fault-tolerance mechanism,” in *Proceedings of the 2017 International Conference on Smart Digital Environment*, pp. 214–219, Rabat, Morocco, July 2017.
- [17] A. Shamir, “How to share a secret,” *Communications of the ACM*, vol. 22, no. 11, pp. 612-613, 1979.
- [18] A. Galletta, L. Bonanno, A. Celesti, S. Marino, P. Bramanti, and M. Villari, “An approach to share MRI data over the Cloud preserving patients’ privacy,” in *Proceedings of the 2017 IEEE Symposium on Computers and Communications (ISCC 2017)*, pp. 94–99, Heraklion, Greece, July 2017.
- [19] E. Alexander and Sathyalakshmi, “Privacy-aware set-valued data publishing on cloud for personal healthcare records,” in *Artificial Intelligence and Evolutionary Computations in Engineering Systems*, pp. 323–334, Springer, Berlin, Germany, 2017.
- [20] L. Sweeney, “Achieving k-anonymity privacy protection using generalization and suppression,” *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, no. 5, pp. 571–588, 2002.
- [21] M. Terrovitis, N. Mamoulis, and P. Kalnis, “Privacy-preserving anonymization of set-valued data,” *Proceedings of the VLDB Endowment*, vol. 1, no. 1, pp. 115–125, 2008.
- [22] M. B. Smithamol and S. Rajeswari, “Hybrid solution for privacy-preserving access control for healthcare data,” *Advances in Electrical and Computer Engineering*, vol. 17, no. 2, pp. 31–38, 2017.
- [23] V. Goyal, O. Pandey, A. Sahai, and B. Waters, “Attributebased encryption for fine-grained access control of encrypted data,” in *Proceedings of the 13th ACM Conference on Computer and Communications Security*, 2006, pp. 89–98, Alexandria, VA, USA, October 2006.
- [24] B. Dhivya, S. P. S. Ibrahim, and R. Kirubakaran, “Hybrid cryptographic access control for cloud based electronic health records systems,” *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, vol. 2, no. 2, 2017.
- [25] K. Shah and V. Prasad, “Security for healthcare data on cloud,” *International Journal on Computer Science and Engineering (IJCSE)*, vol. 9, no. 5, 2017.
- [26] S. Supriya and S. Padaki, “Data security and privacy challenges in adopting solutions for IOT,” in *Proceedings of the 2016 IEEE International Conference on Internet of Cings (iCings) and IEEE green Computing and communications (GreenCom) and IEEE cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData)*, pp. 410–415, Chengdu, China, 2016.
- [27] H. Lohr, A.-R. Sadeghi, and M. Winandy, “Securing the “ e-health cloud,”” in *Proceedings of the ACM international conference on Health informatics—IHI’10*, pp. 220–229, Arlington, VA, USA, November 2010.
- [28] J. L. Griffin, T. Jaeger, R. Perez, R. Sailer, L. Van Doorn, and R. Caceres, “Trusted virtual domains: toward secure distributed services,” in *Proceedings of the 1st IEEE Workshop on Hot Topics in System Dependability (HotDep’05)*, pp. 12–17, Yokohama, Japan, June 2005.
- [29] M. Kumar, M. Fathima, M. Mahendran, 2013, “Personal health data storage protection on cloud using MA-ABE”, *Int J Comput Appl* 2013;75(8):11–6.
- [30] H. Zhu, R. Huang, X. Liu, H. Li, SPEMR: A new secure personal electronic medical record scheme with privilege separation. In: *2014 IEEE International Conference on Communications Workshops (ICC), Sydney, NSW, Australia, 2014*, pp. 700–705.
- [31] M. S. Shin, H. S. Jeon, Y. W. Ju, B. J. Lee, & S. P. Jeong. (2015). Constructing RBAC Based Security Model in u-Healthcare Service Platform. *The Scientific World Journal*, 2015, 1–13. doi:10.1155/2015/937914.
- [32] N. A. Azeez, A. A. Lasisi, 2016. Empirical and statistical evaluation of the effectiveness of four lossless data compression algorithms. *Nigerian J Technol Dev* 2016;13 (2):64–73.
- [33] M. Simplicio, L. Iwaya, B. Barros, T. Carvalho, M. Naslund, 2015, SecourHealth: a delay tolerant security framework for mobile health data collection. *IEEE J Biomed Health Inform* 2015;19(2):761–72.
- [34] M. Barua, R. Lu, X. Liang, X. Shen, 2011. PEACE: An Efficient and Secure Patient Centric Access Control Scheme for eHealth Care System. In: *The First International Workshop on Security in Computers, Networking and Communications, Shanghai, China, 2011*, pp. 970–975.
- [35] L. Guo, C. Zhang, J. Sun, Y. Fang, 2012. PAAS: A Privacy-Preserving Attribute-based Authentication System for eHealth Networks. In: *2012 32nd IEEE International Conference on Distributed Computing Systems,Macau, China, 2012*, pp. 224–233.

# Identification of Retinal Disease using Anchor-Free Modified Faster Region

Arulselvam.T<sup>1</sup>, Dr.S. J. Sathish Aaron Joseph<sup>2</sup>

(Reg No. BDU2120412778809) Research Scholar in Computer Science, J.J college of Arts and Science (Autonomous)

Sivapuram Post, Pudukkottai (Affiliated to Bharathidasan University, Tiruchirapalli), Tamil Nadu, India<sup>1</sup>

Assistant Professor and Research Advisor in Computer Science,(Ref.No:05526/Ph.D.K 10/Dir/Computer Science/R.A) P.G and Department of Computer Science, J.J.College of Arts and Science (Autonomous), Sivapuram, Pudukkottai, Tamil Nadu, India<sup>2</sup>

**Abstract**—Infections of the retinal tissue, as well as delayed or untreated therapy, may result in visual loss. Furthermore, when a large dataset is involved, the diagnosis is prone to inaccuracies. As a consequence, a completely automated model of retinal illness diagnosis is presented to eliminate human input while maintaining high accuracy classification findings. ODALAs (Optimal Deep Assimilation Learning Algorithms) are unable to handle zero errors or covariance or linearity and normalcy. DLTs (Deep Learning Techniques) such as GANs (Generative Adversarial Networks) or CNNs might replace the numerical solution of dynamic systems (Convolution Neural Networks), in order to speed up the runs. With this objective, this study proposes a completely automated multi-class retina disorders prediction system in which pictures from the Fundus image dataset are upgraded using RSWHEs (Recursive Separated Weighted Histogram Equalizations) to boost contrast and noise is eliminated using the Wiener filter. The improved picture is used for segmentation, which is done using clustering and the optimum threshold. The suggested EFFCM is used for clustering (Enriched Fast Fuzzy C Means). The suggested AOO (Adaptive optimum Otsu) threshold technique is used for clustering and picture optimal thresholding. This work suggests AMF-RCNNs (anchor-free modified faster region-based CNNs) that integrate AFRPNs (anchor free regions proposal generation networks) with Improved Fast R-CNNs into single networks for detecting retinal issues accurately. The performances of Accuracy is 98.5%, F-Measure is 96.5%, Precision is 99.2% and different Subset features are 98.5 % show better results when compared with other related techniques or models.

**Keywords**—Retinal disease; fundus image dataset; contrast enhancement; segmentation; Fast Fuzzy C Means; adaptive optimal OTSU; faster region-based convolutional neural network

## I. INTRODUCTION

Human's vision is a crucial sense and lack of which impacts independence and productivities in humans. Retinal illnesses affect millions of humans resulting in visual losses when not identified or treated in early stages of vision loss. DRs (Diabetic retinopathies), glaucoma and other eye defects are macular degenerations which crop up as humans age [1] where early therapies can arrest or eliminate the disease's progressions. Despite the fact that there are many hospitals and eye clinics in India's cities, the doctor-to-patient ratio remains low. There is a paucity of both infrastructure and ophthalmologists in rural locations [2]. Even community outreach activities are hampered by a lack of qualified professionals in remote locations to appropriately assess

patients. With advancements in technology and image processing, it is feasible to automate disease identification and refer patients to doctors for additional evaluation. Ophthalmologists can use retinal fundus imaging to diagnose retinal abnormalities. Early detection can boost treatment chances and avoid blindness [3].

Medical specialists can use retinal fundus images to identify retinal diseases including DRs and retinitis pigmentosa. Studies using MLTs (Machine learning techniques) have been recently focusing on identification of retinal disorders including DRs by categorizing fundus images based on extracted features [5]. The main aspire of this paper is to differentiate automatically abnormalities in the retina without segmentations or feature extractions from their images and use DLTs to automatically categorize retinal images as healthy or sick. Many of these models use the retinal picture to identify, extract, and evaluate disease-specific characteristics. This necessitates a thorough understanding of the illness as well as considerable work in developing the characteristics for the classifiers. Without being taught where to look, MLTs utilize data to uncover hidden patterns and have gained traction in disease diagnostics. These algorithms identify patterns and qualities in images at multiple levels and relate them to recognized disease classes. Guided learning have assisted in early detections and categorizations of eye disorders including cataract, conjunctivitis, and DRs. Various MLTs have been found equivalent to that of human specialists for certain eye conditions in studies. ANNs (Artificial Neuronal Networks) are mathematical models that reproduce neuronal organizations of cerebral cortex [6]. Layers of neurons make up NNs (Neural networks) and these layers are made up of fully connected 'nodes,' each with non-linear 'activation functions,' which reduce errors in executions of GDs (gradient descents) using BPs (back propagations).

Input layers are followed by one or more hidden layers where patterns are processed using system of weights 'connections' for recognizing patterns. These hidden layers get connected to output layers, creating patterns of retinal images. Most ANNs include learning rules which alter weights of neurons in response to input patterns. A typical neural network, on the other hand, is incapable of analyzing patterns in different positions. Specifically, unlike traditional ANNs, the layers of DLTs, such as CNNs and DNNs (Deep Neural Networks) [7], contain neurons organized in three dimensions: width, height, and depth. Certain neurons may sense the

margins while others detect the centre. Using a certain stride and ensuring that various neurons acquire distinct data on pattern localization. Thus, unlike standard NNs, these networks learn more about patterns in images instead of plain detections. Information obtained by first layer's neurons is forwarded to corresponding hidden layers, with neurons contributing vital information about distinct image parts. The last layers use multi-class classification functions that count number of units as output class counts. This research work's contributions include developing DLTs based on CNNs for identifying retinal eye diseases from fundus images and are listed below: RSWHEs are used to boost the contrast of input pictures. Detecting diabetes-related eye disorders areas early

and automatically using cluster-based segmentation is a difficult challenge. The EFFCM-based approach was applied in the described methodology for disease area localization. The findings suggest that combining EFFCM with AOO leads in accurate localization of the afflicted regions, ensuring exact illness detection in an automated way.

The motivation suggested AMF-RCNNs can identify illness symptoms, including early warning indications, and have no difficulty learning to recognize an image of a healthy eye. Finally, we compared the strategy to the most recent cutting-edge methodologies and obtained better performance results.

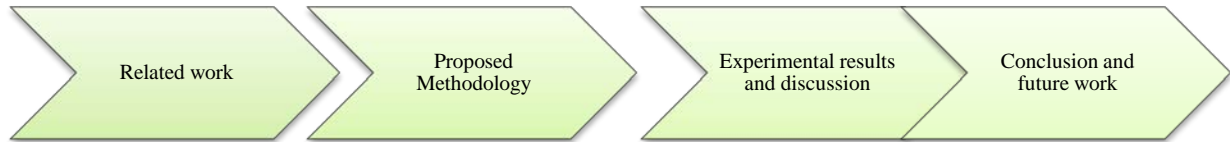


Fig. 1. Examine Method Smart Art.

The overall research technique displayed in Fig. 1, research work summarizes various DLTs techniques applied for detection and classification of eye diseases (Section II). Implementation methodology describes the clustering method and proposed classification method for classifying the retinal disease (Section III). Then the results of the experiment with discussion is projected in experimental and results (Section IV). Finally concludes the paper with future study in conclusion and future work (section V).

## II. RELATED WORK

This section provides a thorough examination of the principles and applications of DLTs in retinal image processing. Machine learning, particularly DLTs, has lately been successfully utilized in this field. This section examines current developments in DLTs approaches for retinal image processing. Pan et al. [8] use DLTs to identify and categorize DRs Lesions in pictures of FFAs (fundus fluoresce in angiographies) by comparing three CNNs: DenseNet, ResNet50, and VGG16. In [9] developed a DLTs-based DC-Gnets (disc cup segmentation glaucoma networks) for glaucoma diagnosis using structural characteristics like cup-to-disc ratios, probability scales of disc damages, inferior/superior nasal temporal regions using RIM-One and Drishti-GS datasets for segmentations. In [10] present a method for detecting the existence of neovascularization that includes image resizes, filtering green channels, Gaussian filters, and morphological methods like erosions and dilations in processing of images. The several layers of CNNs were employed and modelled combined in a VGG-16 net architecture for classification. The method was tested and trained on over 2200 photos from the Kaggle database. In [11] suggested a two-stage approach for detecting and localizing

the optic disc before classifying it as healthy or glaucomatous. The first stage utilizes Regions of CNNs to locate and extract optic discs from fundus images while in the second stage DCNNs (Deep CNNs) classify extracted discs as healthy or glaucomatous. Unfortunately, none of the publicly obtainable retinal fundus image datasets could provide much needed bounding boxes localization of discs.

In [12] offered modified U-Net topologies that included block squeeze and excitations as well as usages of attention gates in datasets to segment demarcation lines/ridges and vessels. These changes were verified and confirm by ROP experts. All three networks (AG U-Net, U-Net and SE U-Net) showed good results with ranges in the dice coefficients ranging from 1 to 6% for the HVDROPDB datasets. In [13] suggested architecture employs a dilated convolution filters to obtain bigger receptive fields, resulting in near-human accuracy in segmenting retinal blood vessels. The popular datasets were used to train the CNNs. In [13] introduced the DL-CAEF (DLT Assisted Convolution Auto-Encoders Frameworks) to diagnose glaucoma and recognize AVPs (Anterior Visual Pathways) from retinal fundus pictures. In [15] identified CWS, HE on retinal images using super iterative clustering approach which comprised of CNNs and encoders with encoder structures. To convert red, blue and green images into ideal grayscale images devoid of noises, FBMIR dataset's data were combined with histogram filters and subsequently categorized using DALAs (deep assimilation learning algorithms). Through the investigation of retinal fundus color photographs, in [16] proposed presenting a mixture of losses in DNNs model to enhance recognition performances in biomedical data for classifiers. Table I show the literature review method with advantages and disadvantages.



TABLE I. THE LITERATURE REVIEW METHOD WITH ADVANTAGES AND DISADVANTAGES

Authors	Method	Advantages	Disadvantages	Evaluation result
Juneja et al., [9-8]	DC-Gnet	Reducing the computational complexity.	It is susceptible to interference from uneven lighting in fundus pictures, resulting in low precision and poor robustness.	Dice, Jaccard, and recall coefficients
Bajwa et al., [10-11]	Deep CNNs	This strategy avoids the necessity for the creation of dataset-specific empirical or heuristic localizations.	However, if glaucoma is detected early enough, it is feasible to delay the progression of the disease.	95% accuracy
Agrawal et al., [12]	U-Net architectures	As a result, trained models were generic and robust to data fluctuations or heterogeneities.	However, U-net-based segmented discs are extremely sensitive to segmentation accuracy, and even little errors in eye disease delineation can have a major impact on diagnosis.	AUC obtained for all three networks was above 0.94, AG U-Net sensitivity of 96% and specificity of 89%
Biswas et al., [13]	CNNs	Obtaining appropriate contexts for items extending beyond filter sizes	However, because to the convolutional filters' restricted receptive field, this approach frequently fails to effectively segment the retinal blood vessels and introduces extra noise.	AUC) of 0.9794 and an accuracy of 95.61%
Saravanan et al.,[14]	DL-CAEF and CNNs	It used multi-model learning for reducing image reconstructions or classification errors.	However, if ambient light enters the picture as it is being captured, it might appear brighter than the optic disc.	IOU greater than 50%
Sikkandar [15]	DALA	The ability to learn discriminative depiction of eye illness performed much better.	However, there may be additional bright areas in the picture owing to illness or poor image capture settings, which can impair the DALA method's efficacy.	accuracy ratio of 98.5%

### III. PROPOSED METHODOLOGY

The suggested AMF-RCNN for classifying retinal fundus pictures was examined. BDR Backgrounds for PDR stands for Proliferative DRs, CRVOs stand for Central Retinal.

Vein Occlusions, CNVs stand for Choroidal Neovascularizations, HISTs stands for Histoplasmoses, and Normal stand for Normal eyes. Fig. 2(a) depicts the normal anatomical features of the retina, whereas Fig. 2(b) depicts an eye illness complication in a retina.

In Fig. 3 depicts the broad framework of the primary sequential procedures for classifying DR pictures in the suggested technique. To begin, DR pictures were acquired,

graded, and tagged. The STARE (Structured Analysis of the Retina) web-database was used to obtain the dataset. The collection contains 400 raw fundusoscopic pictures from various instances, including 13 illnesses and normal cases [17]. After that, the appropriate image pre-processing techniques were used to capture images for better training for the planned learning. The improved picture is used for segmentation, which is done using clustering and the optimum threshold. The suggested method gathered features from multiple sample photos with ground truth labels and automatically changed its hyper parameters to get the greatest classification accuracy. The identification outcome might help ophthalmologists decide whether or not to refer patients.

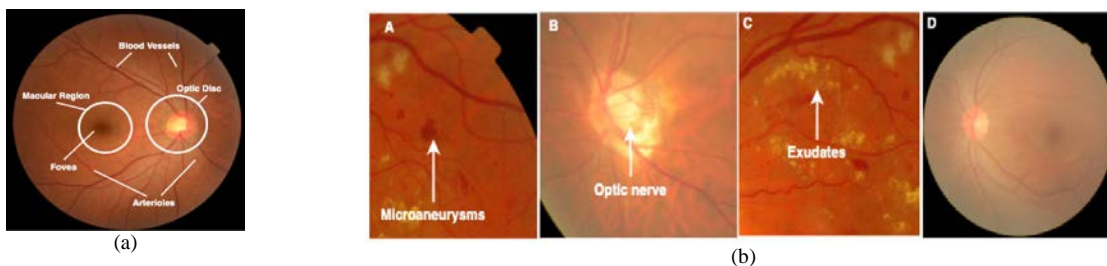


Fig. 2. (a) The Normal Anatomical Structures of the Retina. (b) Illustrates a Complication of Eye Disease in a Retina: A. Microaneurysms, Narrow Bulges (DRs), B. Optic Nerve Damage (Glaucoma), C. Exudates with Retinal Thickening (Diabetic Macular Edema), D. Degeneration of Lens (Cataract) [4].

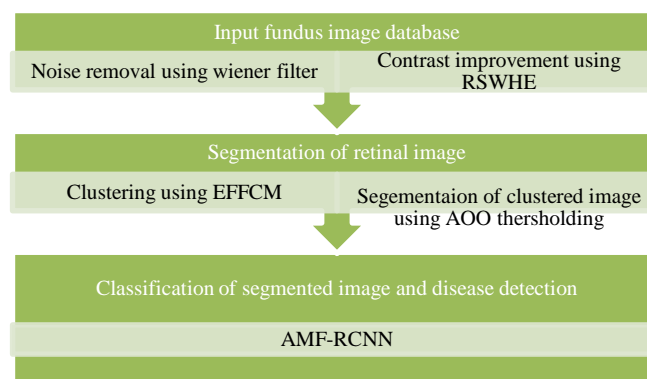


Fig. 3. The General Framework of Proposed Methodology.

### A. Image pre-processing

The suggested approaches aid in the identification of eye diseases by enhancing picture quality and clarity. Here, two ways for improving the contrast and quality of retina fundus pictures have been developed: RSWHEs boost picture contrast, whereas Wiener filters reduce noise. RSWHEs: RSWHEs work by recursively segment input histograms into sub-histograms and modify sub-histograms with weights based normalized power law function procedures and histogram equalizations on weighted sub-histograms separately. It improves the contrast of a picture in three phases [18]: First, take the picture  $I$  and calculate the histogram  $H(I)$ , which is then split into the number of sub-histograms. A second histogram weighting is used to modify the sub-histograms using a normalized power law. Finally, histogram equalization is performed, in which the weighted sub-histograms are equalized separately over the modified sub-histograms. Wiener filters: Wiener filters are non-adaptive additional predefined patterns that incorporate linear approximations of predicted signal organization on or after a non-adaptive extra predetermined pattern. A typical Wiener filter is a convolution filter that computes the shape and size of the neighborhood using only a frame. It may be more effective to filter with more really large Viennese masks and fewer blurry masks. A fundus photograph of size  $(ij)$  must have each pixel representing the strength of a single stationary point in front of the camera. The Wiener filter is used to remove noise from a signal that has been tampered with by statistical analysis. The Wiener filter is a type of filter that is used to reduce assuming that signal and noise processes are second-order picture derivatives. Consider the following equation-1 to demonstrate and transform to 2D:

$$ob(i, j) = ir(i, j) * I(i, j) + un(i, j) \quad (1)$$

Where  $*$  represents  $I$  is the unknown real ovarian cyst picture with  $ij$  pixel value,  $ir$  is the impulse response of a linear, time-invariant filter,  $un$  is incremental unknown noise independent of  $I$ , and  $ob$  is the observed image. Then, as given in equation-2, a de-convolution filter  $dc$  must be found to evaluate  $I$ :

$$\hat{I}(i, j) = dc(i, j) * o(i, j) \quad (2)$$

where  $\hat{I}$  indicates the value of  $I$  that reduces the mean square error. The transition function of  $Tr$ , in the frequency domain, given in equation-3,

$$Tr(x, y) = \frac{ir^*(x, y)PS(x, y)}{|ir(x, y)|^2PS(x, y) + NSP(x, y)} \quad (3)$$

$Tr(x, y)$  is the Fourier transform of the probability mass equation,  $PS$  is the power spectrum of the signal process used to acquire the Fourier transform of the signal collinearity, and  $NSP(u, v)$  is the power spectrum of the noise ( $N$ ) practise acquire by generating the Fourier transform of the noise autocorrelation.

### B. Segmentation Process using EFFCM Clustering and AOO Method

The improved picture is used for segmentation, which is done using clustering and the optimum threshold. The suggested EFFCMC Algorithm is used for clustering, while the proposed AAO threshold with ALOs is used for optimum thresholding (Ant Lion Optimizations). Real-time photos are used to collect materials for processing. FCMC is a widely used approach for picture segmentation since it is more efficient than other machine learning algorithms. The biggest disadvantage of this approach was its slowness. The Fast Fuzzy C Means Clustering technique was applied to increase the speed of this technique. The primary distinction in this approach was that an image histogram was employed instead of raw picture pixels. The objective function OF of the Fast Fuzzy C Means Clustering method is given by,

$$OF = \sum_{i=0}^n \sum_{c=1}^c hist_i * \mu_{ic} * dis(i, \theta_c)$$

Here,  $hist_i$  refers to the histogram,  $\mu_{ic}$  refers to the fuzzy membership between pixel  $x_i$  and histogram of cluster with center  $\theta_c$ ,  $dis(i, \theta_c)$  refers to the distance between pixel  $x_i, i = 1, 2, \dots, n, \forall n = 255$  and histogram of cluster with center  $\theta_c$ , Let  $C = 1, 2, \dots, c$  represent each centroid. Then, the mean of the pixels in every centroid is given as  $v_c$ . The mahalanobis distance between the data point  $x_i$  with each  $v_c$  is then calculate. This term is added to the objective function in the present EFFCMC plan. Thus, the new objective function is given as

$$OF = \sum_{i=0}^n \sum_{c=1}^c hist_i * \mu_{ic} * dis(i, v_c) * mahalan(x_i, v_c)$$

Otsu thresholds are common image segmentation techniques that use global thresholds to divide images into two categories: foreground and background. Pixel values in images are compared using thresholds. When values of pixels exceed threshold values they are classified as foregrounds; else they are classified as backgrounds. The initial threshold determines Otsu algorithm's efficacy. As a result, the authors suggest an adaptive Otsu thresholding method (AOTA) in which optimal threshold selection is performed using ALOs that adaptively select best starting threshold values. In this part, the variable threshold influences the AOTA output, and the goal of any optimizer is to find values for these variables that provide the best segmentation rate and accuracy. The Ant Lion Optimizer (AOO) is used to increase the performance of AOTA. AOO is a strategy that disallows the local minima and maxima in order to obtain an optimal solution. To compute the input for the AOO method, the ALO [17] algorithm is employed. The suggested algorithm AOTA system is adjusted utilizing the ALO optimizer using the provided fundus image pre-processed data set. This raises the bar for learning from the model. ALO is used to modify the incidence of local minima and maxima. The AOTAs through ALOs is illustrated in Fig. 4.

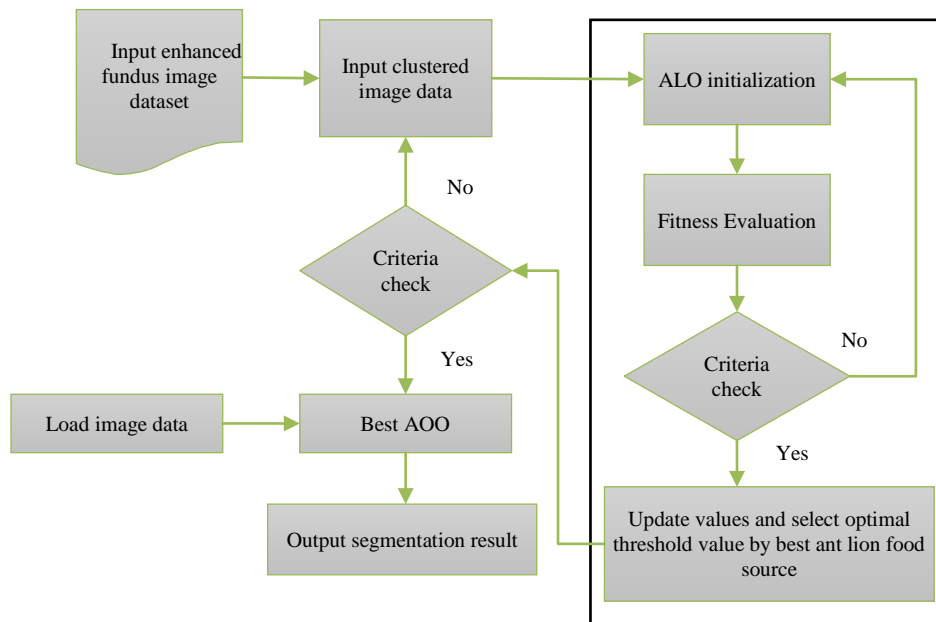


Fig. 4. Flowchart of Proposed AOO Method.

Early data quality segmentation is a necessary step since classifying early phases of eye illnesses includes uncertainties that govern disease predictions, and progressions caused by aspects of image data which necessitate strong models to reliably predict outcomes and the segmentation region of eye disease. In this study, AOO is used for segmentation with the weaker domain identification among the major four domains.

- Step 1. Initialize: random initialization of the positions of input variables of the features as Ant lion selected from preprocessed image dataset.
- Step 2. Calculate the cumulative total of a maximum number of iterations, where iteration represents the steps in a random walk. One matrix stores the position of each input's value. Another matrix stores the relevant objective values. Another matrix is generated in order to save the position and fitness value as accuracy.
- Step 3. Use a random threshold value to update the location of the input value.
- Step 4. Create Traps: Make two vectors, one with a minimum of all variables from a single input source and the other with a maximum of all variables from the same input source. This provides the input weight for the fitter for the intended output value.
- Step 5. Entrapment of Ants in Traps: If it gets fitter, replace the position of all input variables with the appropriate fit of the other input variables.

- Step 6. Finally, change the threshold value based on Ant Lion's location.
- Step 7. Rebuilding Traps: Check the termination requirements; if the termination conditions are met, return the ideal solution; otherwise, proceed to updating Ant Lion's location.

Following inference is used to provide automated threshold selection for segmentation. Elitism, or remembering the best solution discovered, is an important feature of a nature-inspired algorithm that allows for the preservation of the best solution achieved at any point of the optimization process. The best result generated in each iteration is kept and considered Elite in this investigation. Because the Elite is the fittest output, it will influence the movements of all other variable thresholds throughout iteration.

### C. Eye Disease Image Detection Model using AMF-RCNN Model

As illustrated in Fig. 5, the proposed eye disease image detection model, AMF-RCNN, consists of eye disease images. Area of interest extraction network, a features extraction network, and an eye illness picture detection network are all examples of networks. To begin, CNNs are used to extract characteristics and get fusion features from an image of an eye ailment. Second, the AFRPN receives the fused feature map produce by the feature fusion methods and creates a series of eye sickness image suggestions. The top sorted are reduced to the similar size ROI vector using the ROI Align layer [12]. The classification layer predicts the image category of an eye condition, and the bounding box regression layer refines class-specific suggestion box offsets, using these fixed vectors.

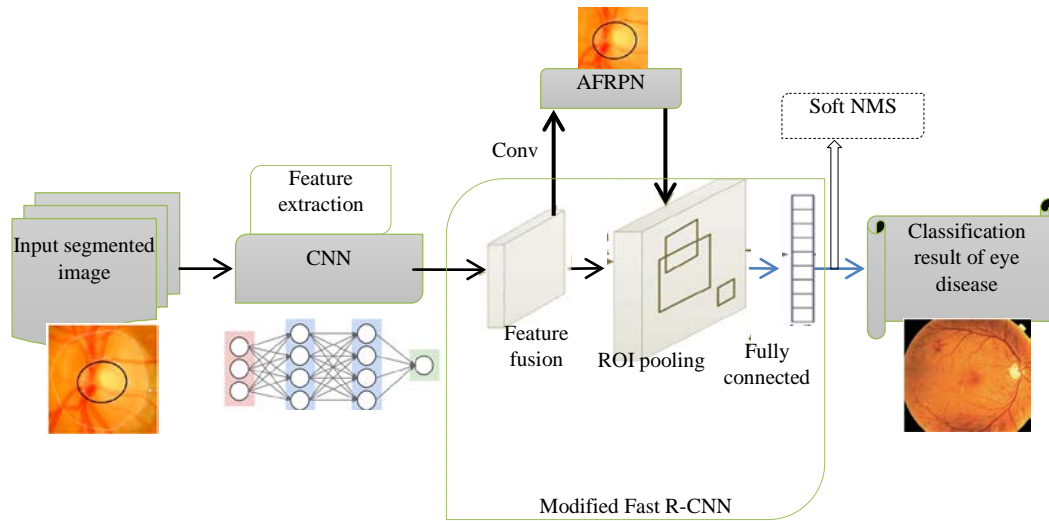


Fig. 5. Block Diagram of AMF-RCNN.

**Feature Extraction Network:** Automatically extracting target features from eye disease pictures and integrating feature depiction and goal for "joint learning" were accomplished using the CNN-based approach. Depending on the optimization goal, the parameters of eye sickness picture features can be changed adaptively during network training. The feature extraction network is made up of four layers: convolution, pooling, activation, and feature fusion. In conclusion, the new feature fusion module creates a more compact feature map with strong semantic characteristics and fine-grained details ideal for micro item identification. AFRPNs are full convolution networks that enable compute sharing with multi-class detection networks. This tiny network employs a 33 spatial window on the input feature map, and the feature extract by every sliding window is translated to a 512-dimensional feature vector using a 33kernel convolutional operation with 512 channels, which is then transmitted to two simultaneous complete convolutional networks. For classification, one branch generates two probabilities of being objects or not, while another generates four box coordinates for localizations.

This study integrates AFRPN and modified Fast R-CNNs into one network, i.e. AMF-RCNN, to identify the multi-categories eye illness picture, where AMF-RCNN is trained end-to-end via back-propagation and stochastic GDs (SGD), allowing for shared convolutional layers. Each SGD mini-batch is created for AMF-RPN training from a single eye illness imaging image containing 256 data. Positive and negative samples are chosen at random for each mini-batch, so that the ratio of positive to negative samples is 1:1. The regression layers and classification start with a zero-mean Gaussian distribution with standard deviations of 0.01. Backward propagation occurs as normal, and the backward propagated signals for the shared convolutional layer originate from a mix of AFRPN loss and modified Fast R-CNNs loss. The fundus image scale varies substantially during the real eye disease detection procedure. The original quicker RCNN typically uses a set size for all training pictures. As a result, the generalization performance of object identification with varying sizes is low. This work employs multi-scale training;

before being uploaded to the network, the photographs will be scaled at random. After that, the various scale photographs will be taught. For joint training, loss function LF is widely used. Furthermore, enhancing localization exactness can get better the overall finding rate of airports. As a result, we get better the classification loss, which is defined as the integral of the classification loss  $Loss_{cl}$  under IoU threshold  $th$ . The final loss function is:

$$\int_{50}^{100} Loss_{cl}(pr_{th}, clp_{th}) \approx \sum_{50}^{100} \frac{Loss_{cl}(pr_{th}, clp_{th})}{n}$$

Where  $n = 4$ ,  $th \in \{50,60,70,80\}$ , the predicted probability of the final output of the network model is the average value of different thresholds. The feature map corresponding to the suggestion box is input to the full connection layer, and finally, output from two same output layers. One output is the set of predicted probabilities for the background and  $k$  classes objects, namely,  $(pr_0, pr_1 \dots pr_k)$ ,  $clp_{th}$  represents the class prediction value when the threshold value is  $th$ .  $pr_{th}$  Represents the prediction probability and  $n$  is the equal number in definite integral interval. The final loss function can be rewritten as:

$$LF(pr, clp, v) = Loss_{cl}(pr, clp) + [clp \geq 1] \frac{Loss_{cl}(pr_{th}, clp_{th})}{n}$$

When many types of eye illness images are detected during the testing stage, duplicate boxes are frequently predicted for the same eye disease image. Traditionally, a recommendation box SB with the highest score is chosen, but its adjacent boxes are suppressed depending based on a preset overlap threshold. If the suppressed boxes include extra items, however, the suppression will cause some items to be overlooked. To solve this problem, instead of forcing class scores to be zero, give them penalty weight. The following is an example of how a Gaussian penalty function (GPF) is used:

$$GPF = cs_i e^{-\frac{IoU(SB, nb_i)^2}{th}}$$

$$IoU(SB, nb_i) = \frac{area(SB \cap nb_i)}{area(SB \cup nb_i)}$$

Where  $cs_i$  is the confidence score for the  $i$ -th detection box,  $IoU(SB, nb_i)$  denotes IoU (Intersection over Unit) ratio between boxes with highest scores  $SB$  and their neighboring boxes  $nb_i$  and  $th$  is threshold is set to 0.5. Following the above stages, sharper eye disease image identification results will be obtained. When we find multi-category pest items during the testing stage, duplicate boxes are frequently anticipated for the same bug. To suppress these redundant boxes, a class specific NMS is traditionally used: picking a bounding box with the highest score but suppressing its nearby boxes depending on a pre-defined overlap level (Dalal and Triggs, 2005). However, if the suppressed boxes include additional items, the suppression will result in some things being missed. To remedy this issue, we provide penalty weight to class scores instead of forcing them to be zero. A Gaussian penalty function is used specifically:

$$cs_i = e^{-\frac{IoU(SB, nb_i)^2}{\delta}}$$

Where  $\delta$  is set to 0.5 and after the above soft-NMS suppression method, can obtain finer eye disease detection results.

#### IV. RESULTS AND DISCUSSION

To validate the effectiveness of the proposed approach, the proposed AMF-RCNN method is used to identify eye diseases and is compared to several current detection methods based on CNNs [8, DL-CAEF [14], and ODALAs. The fundus pictures were arbitrarily divided into training (70%) and test (30%) (30%). There were 650 training photos and 280 validation images from five illness categories used. The results showed that the outputs provided by the proposed single CNNs model, which was meant to categorize pairs of distinct retina diseases BDRs (Background DRs), CRVOs CNVs, HISTs: PDRs: Proliferative DRs and Normal were accurate, sensitivity, and specificity. In the context of classification presentation, a true positive value (TP)" is a result that properly predicts the positive class. A true negative value (TN) is a result that forecasts the negative class appropriately. False negative (FN) and false positive (FP) values are used to reflect misclassified samples. Using accuracies, precisions, F1-scores, and MCCs as parameters (Mathew Correlation-coefficients), specificities, and sensitivities, their equations may be represented as follows:

Accuracies of models show overall performances of models and computed by the formula given below:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} * 100$$

Precisions are ratios of actual positive scores and positive predicted scores by classification models/algorithms. Precisions are computed by the following formula:

$$Precision = \frac{TP}{TP + FP} * 100$$

F1-scores are weighted measures of both recalls and precisions. They range between 0 and 1 where 1 implies good performances of classification algorithms while 0 stands for bad performances.

$$F1\ score = \frac{2 * Precision * Recall}{Precision + Recall}$$

MCCs are correlation coefficients between actual and predicted results. MCCs give resultant values between -1 and +1. Where -1 represents completely wrong predictions of classifiers while 0 implies classifiers generate random predictions and +1 represents ideal predictions of classification models. The formula for calculating MCCs are given below:

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

Specificities are ratios of recently classified healthy people to total counts of healthy people. It implies the predictions are negative and persons are healthy. The formula for calculating specificity is given as follows:

$$Specificity = \frac{TN}{TN + FP} * 100$$

Sensitivity is the ratio of recently classified heart patients to the total patients having heart disease. It means the model prediction is positive and the person has heart disease. The formula for calculating sensitivity is given below:

$$Sensitivity = \frac{TP}{TP + FN} * 100$$

From Fig. 6, it gives the accuracy of proposed and existing models for the amount of features in a given database. The AMF-RCNN increases the accuracy while reducing the processing time. The AMF-RCNN attains the accuracy of 98.5% compared to all other models since the high quality generated by the proposed AFRPN can be applied to improve the disease detection. The results of existing methods such as DL-CAEF, CNNs and ODALAs are 94%, 96% and 98% respectively. Thus the proposed algorithm is greater to the existing algorithms in terms of better good validation results for predicting DR disease.

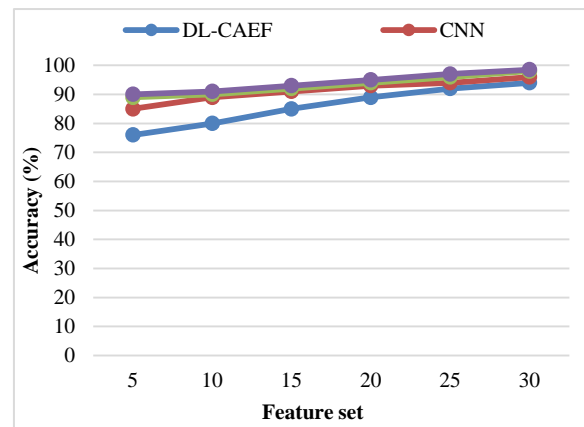


Fig. 6. Accuracy Performance of Methods on different Subsets of Feature.



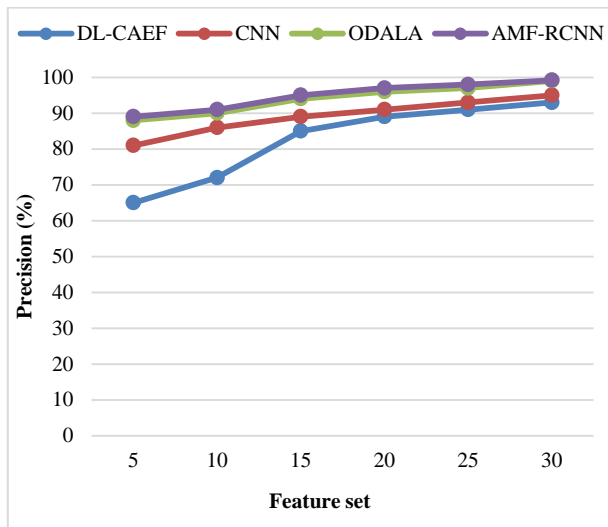


Fig. 7. Precision Performance of Methods on different Subsets of Feature.

From Fig. 7, it indicates the precision of proposed and existing models for the amount of features in a given database. As increasing the amount of features, the precision is also maximized. The AMF-RCNN attains a recall of 99.2%. This may be due to the benefit of our sample selection approach, which is based on an effective region of RFs and allows more tiny objects to participate in training the eye illness detection module. Existing approaches such as DL-CAEF, CNNs, and ODALAs achieve 93 per cent, 95 per cent, and 99 per cent accuracy, respectively. Existing approaches cannot yield good detection findings when blood vessels are thickly scattered. In comparison to these three approaches, the suggested method is capable of successfully overcoming blood vessel interference and obtaining the best eye disease detection result.

From Fig. 8, it indicates that the F1-score of proposed and existing models for the amount of features in given databases. While maximizing the amount of features, the f-measure is also maximized. For e.g., the AMF-RCNN provides an f-measure of 96.5% compared to the all other models such as DL-CAEF, CNNs and ODALAs. The results of existing methods such as DL-CAEF, CNNs and ODALAs are 91%, 95% and 96% respectively. The DL-CAEF approach clearly fails to capture the limits of the eye illness adequately, resulting in a fitted ellipse that deviates significantly from the ground truth. The CNNs approach is sensitive to weak edges and produces the lowest results for detecting eye diseases. The ODALAs approach detects the brightest section of the eye illness zone, resulting in incorrect detection findings. Instead, the suggested technique, which takes use of the deep features recovered from the pretrained AFRPN, is a smaller amount influenced by blurred eye disease borders and poor contrast and achieves the desired eye illness identification results.

From Fig. 9, it indicates the MCC of proposed and existing models for the amount of features in a given database. As increasing the amount of features, the MCC is also maximized. For e.g., the AMF-RCNN attains a recall of 98.5% compared to the DL-CAEF, CNNs and ODALAs. The performance of existing approaches is as follows, based on the examination of experimental findings. The available

approaches are substantially hampered by the interference of lesions. When the intensities of the eye illness region and lesions are near, the system cannot identify them accurately. In comparison to the preceding techniques, the suggested method may resist the impact of lesion interference to a assured amount and take out the eye disease borders more precisely using an efficient clustering method.

From Fig. 10, it indicates the recall of proposed and existing models for the amount of features in a given database. As increasing the amount of features, the recall is also maximized. For e.g., the AMF-RCNN attains a recall of 98% compared to the DL-CAEF, CNNs and ODALAs. By observing the results, the proposed method has attain alike results and is superior than the other methods, because of effective EFFCM and adaptive OTSU threshold with ALO segmentation method could achieve prior eye disease detection.

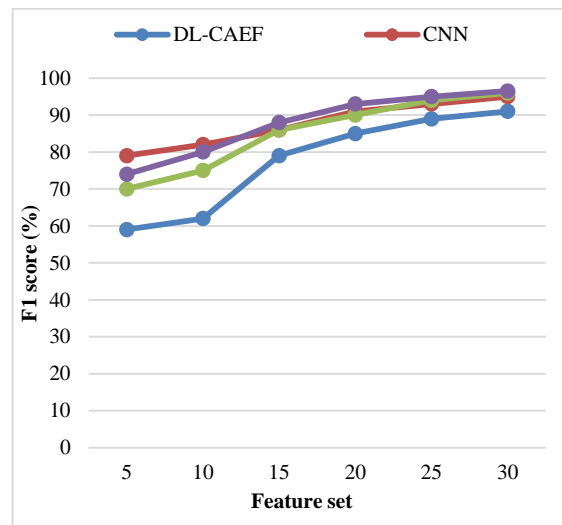


Fig. 8. F1-score Performance of Methods on different Subsets of Feature.

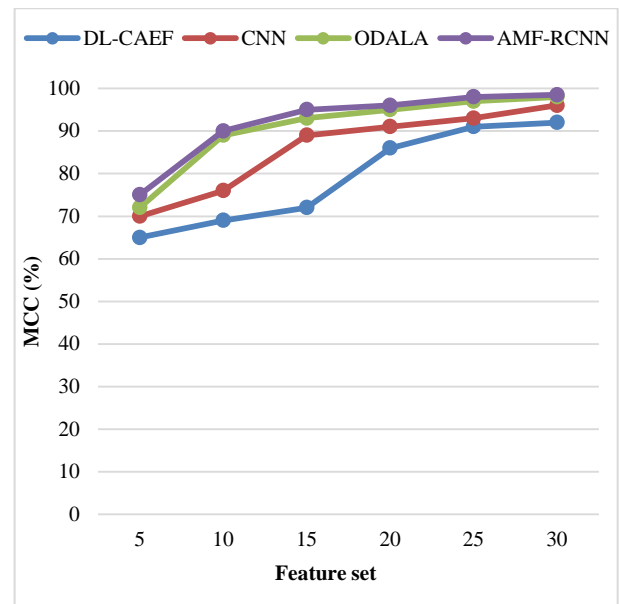


Fig. 9. MCC Performance of Methods on different Subsets of Feature.



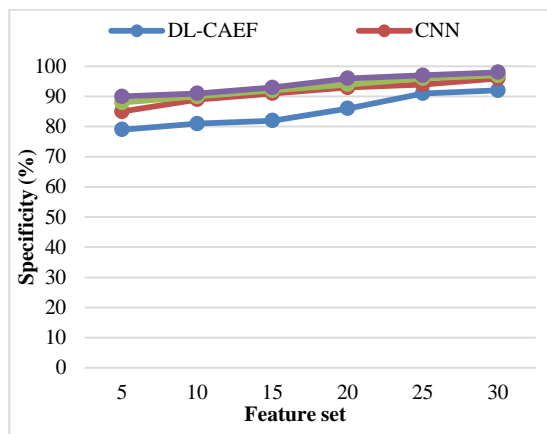


Fig. 10. Specificity Performance of Methods on different Subsets of Feature.

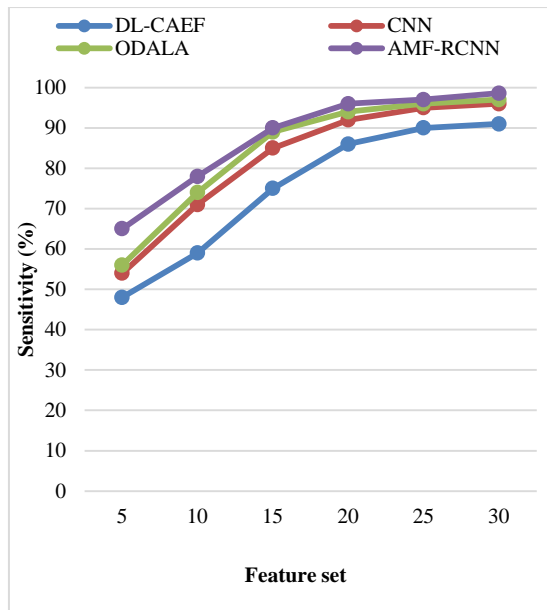


Fig. 11. Sensitivity Performance of Methods on different Subsets of Feature.

From Fig. 11, it gives the accuracy of proposed and existing models for the amount of features in a given database. The AMF-RCNN increases the accuracy and attains the accuracy of 98% compared to DL-CAEF, CNNs and ODALAs. Thus the proposed algorithm is greater to the existing algorithms in terms of better good validation results for predicting DR disease. Because this proposed method tends to focus on detection efficiency with including the detection accuracy of targets, also benefiting by the feature fusion module.

## V. CONCLUSION AND FUTURE WORK

In this work, we describe a novel unsupervised learning strategy for detecting eye diseases in retinal fundus pictures using AMF-RCNN and EFFCM-AOO segmentation. In addition, pay attention to the precise extraction of the illness region in the presence of vascular structures, lesion regions, and intensity in homogeneity. The EFFCM-AOO-based segmentation detection approach is used first to pre localize the eye disease region. On this premise, the AMF-RCNN model is developed to extract the correct optic disc area by

including the deep features generated from pre-trained AFRPNs into the original framework. The experimental findings and quantitative analysis show that the suggested technique is capable of precisely detecting eye disease areas and outperforms several current methods. AMF-RCNNs have shown promising results in the identification of eye diseases, but the complexity of regulating is unknown and regarded a black box. In the future, for example, research should focus on fine-tuning the restrictions of the existing AMF-RCNN approach in order to improve classification efficiency. As a result, determining the efficient model and ideal values for the number of hidden layers and modules in various levels remains difficult.

## REFERENCES

- [1] Ouyang, Y., Heussen, F. M., Keane, P. A., Sadda, S. R., & Walsh, A. C. (2013). The retinal disease screening study: prospective comparison of nonmydriatic fundus photography and optical coherence tomography for detection of retinal irregularities. *Investigative ophthalmology & visual science*, 54(2), 1460-1468.
- [2] Khan, R., Surya, J., Rajalakshmi, R., Rani, P. K., Anantharaman, G., Gopalakrishnan, M., ...& Raman, R. (2021). Need for Vitreous Surgeries in Proliferative Diabetic Retinopathy in 10-Year Follow-Up: India Retinal Disease Study Group Report No. 2. *Ophthalmic Research*, 64(3), 432-439.
- [3] Prashantha, G. R., &Patil, C. M. (2018). An approach for the early detection of retinal disorders and performing human authentication. In *Proceedings of International Conference on Cognition and Recognition* (pp. 157-173). Springer, Singapore.
- [4] Sarki, R., Ahmed, K., Wang, H., & Zhang, Y. (2020). Automatic detection of diabetic eye disease through deep learning using fundus images: a survey. *IEEE Access*, 8, 151133-151149.
- [5] Gnanaselvi, J. A., &Kalavathy, G. M. (2021). Detecting disorders in retinal images using machine learning techniques. *Journal of Ambient Intelligence and Humanized Computing*, 12(5), 4593-4602.
- [6] Harun, N. H., Yusof, Y., Hassan, F., &Embong, Z. (2019, April). Classification of fundus images for diabetic retinopathy using artificial neural network. In *2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT)* (pp. 498-501). IEEE.
- [7] Deng, L., & Yu, D. (2014). Deep learning: methods and applications. *Foundations and trends in signal processing*, 7(3-4), 197-387.
- [8] Pan, X., Jin, K., Cao, J., Liu, Z., Wu, J., You, K., ...& Ye, J. (2020). Multi-label classification of retinal lesions in diabetic retinopathy for automatic analysis of fundus fluorescein angiography based on deep learning. *Graefes' Archive for Clinical and Experimental Ophthalmology*, 258(4), 779-785.
- [9] Juneja, M., Thakur, S., Wani, A., Uniyal, A., Thakur, N., & Jindal, P. (2020). DC-Gnet for detection of glaucoma in retinal fundus imaging. *Machine Vision and Applications*, 31 (5), 1-14.
- [10] Saranya, P., Prabakaran, S., Kumar, R., & Das, E. (2022). Blood vessel segmentation in retinal fundus images for proliferative diabetic retinopathy screening using deep learning. *The Visual Computer*, 38(3), 977-992.
- [11] Bajwa, M. N., Malik, M. I., Siddiqui, S. A., Dengel, A., Shafait, F., Neumeier, W., & Ahmed, S. (2019). Two-stage framework for optic disc localization and glaucoma classification in retinal fundus images using deep learning. *BMC medical informatics and decision making*, 19(1), 1-16.
- [12] Agrawal, R., Kulkarni, S., Walambe, R., Deshpande, M., &Kotecha, K. (2022). Deep dive in retinal fundus image segmentation using deep learning for retinopathy of prematurity. *Multimedia Tools and Applications*, 81(8), 11441-11460.
- [13] Biswas, R., Vasan, A., & Roy, S. S. (2020). Dilated deep neural network for segmentation of retinal blood vessels in fundus images. *Iranian Journal of Science and Technology, Transactions of Electrical Engineering*, 44(1), 505-518.

- [14] Saravanan, V., Samuel, R., Krishnamoorthy, S., & Manickam, A. (2022). Deep learning assisted convolutional auto-encoders framework for glaucoma detection and anterior visual pathway recognition from retinal fundus images. *Journal of Ambient Intelligence and Humanized Computing*, 1-11.
- [15] Sikkandar, M. Y. (2021). Automatic Detection of Genetics and Genomics of Eye Disease Using Deep Assimilation Learning Algorithm. *Interdisciplinary Sciences: Computational Life Sciences*, 13(2), 286-298.
- [16] Luo, X., Li, J., Chen, M., Yang, X., & Li, X. (2021). Ophthalmic disease detection via deep learning with a novel mixture loss function. *IEEE Journal of Biomedical and Health Informatics*, 25(9), 3332-3339.
- [17] Zago, G. T., Andreão, R. V., Dorizzi, B., & Salles, E. O. T. (2020). Diabetic retinopathy detection using red lesion localization and convolutional neural networks. *Computers in biology and medicine*, 116, 103537.
- [18] Kim, M., & Chung, M. G. (2008). Recursively separated and weighted histogram equalization for brightness preservation and contrast enhancement. *IEEE Transactions on Consumer Electronics*, 54(3), 1389-1397.

# OpenCV Implementation of Grid-based Vertical Safe Landing for UAV using YOLOv5

Hrusna Chakri Shadakshri V<sup>1\*</sup>  
Electronics & Communication  
Engineering  
BMS College of Engineering  
Bangalore, India

Veena M. B<sup>2</sup>  
Senior IEEE Member  
Electronics & Communication  
Engineering  
BMS College of Engineering  
Bangalore, India

Keshihaa Rudra Gana Dev V<sup>3</sup>  
AXG Group  
Intel Technology India Pvt Ltd  
Bangalore, India

**Abstract**—The challenge of proving autonomous landing in practical situations is difficult and highly risky. Adopting autonomous landing algorithms substantially minimizes the probability of human-involved mishaps, which may enable the use of drones in populated metropolitan areas to their full potential. This paper proposes an Unmanned Aerial Vehicles (UAV) vertical safe landing & navigation pipeline that relies on lightweight computer vision modules, able to execute on the limited computational resources on-board a typical UAV. In this work, a grid-based mask technique is proposed for selecting the safe landing zones where each grid is parameterizable based on the size of the UAVs, which is implemented using OpenCV. A custom trained YOLOv5 model is the underlying building block for safe landing algorithm which is trained for aerial views of pedestrians, cars & bikes to identify as obstacles. The nearest obstacle-free zone algorithm is applied over the YOLOv5 output where boundary box locations are identified using Hue Saturation Value (HSV) filtering and then split into grids for safe landing zones where maximum coverage is taken into account while analyzing each scene. It performs a 2-level operation to prevent collisions while descending at different altitudes. Since UAV is expected to be processing only at predetermined altitudes, which will shorten the processing time, generating a PID signal for UAV actuators to navigate to the required safe zone with utmost safety and accuracy.

**Keywords**—Autonomous UAV system; computer vision algorithm; YOLOv5; safe landing site selection; Haversine equations

## I. INTRODUCTION

Unmanned Aerial Vehicles (UAVs) is that uses navigation and control software powered by artificial intelligence (AI) and do not need a human pilot to fly them. These aircraft carry out activities and make decisions on their own, from takeoff and landing to conducting aerial site inspections and surveys. The utility of an autonomous UAV hinges on its ability to navigate with acceptable positional inaccuracy. The term "autonomous UAV" means a UAV that can fly without external guidance with help of onboard sensors and processors. The main problem is to build UAVs strong enough to fly independently and land safely in open fields without harming people, as in automated package delivery applications [17].

The law prohibits UAV operation over a crowd, but in practice, UAVs may fly over an unwary crowd, compromising

people's safety in the event of a failure, such as a communication loss, a power shortage, or a human error. Providing every UAV with safety rules to prevent injuring people in an emergency and to select landing zones autonomously is vital. If every UAV had an emergency autonomous landing system [15], it would minimize harming people during drone accidents and enhance their urban deployment potential, especially in crowded circumstances.

Different drones use obstacle avoidance sensors such as stereo vision, ultrasonic (Sonar), time-of-flight, lidar, infrared, and monocular vision, either singly or in combination, thus fusing data for complex computations [18]. The data from these numerous obstacle avoidance sensors is fed back to the flight controller, which further uses algorithms and software to detect obstacles. The role of the flight controller is diverse, one of which is the real-time processing of visual data from the environment that is scanned by the obstacle detection sensors that will be employed in our model.

UAVs' limited processing capability owing to weight and battery power limits is one major challenge and developing UAV-compatible algorithms is difficult too. Also, Deep Neural Networks (DNNs) attaining state-of-the-art computer vision results demands a lot of processing resources for real-time operation. Utilization of single-stage object detection algorithm [8] like YOLOv5 along with OpenCV functions in the proposed architecture aids in surpassing the above concerns thus maximizing the desired outcome with minimal resources.

This paper is organized as follows. Section II described the related work for this study. In Section III, a proposed architecture for autonomous UAV safe landing algorithm is discussed. Simulation results are described in Section IV and Section V concludes the paper.

## II. RELATED WORK

Human recognition based on live input is crucial to the safety of autonomous UAV flying [1]. In order to avoid hurting people in the event of a failure, UAV safe landing demands that the UAV visually detect people [2] nearby the landing place; airspace above/near humans should be treated as a no-fly zone. Such detectors place a properly sized rectangular bounding box around each item they find on the image feed and assign it a discrete class label.

\*Corresponding Author.

Early deep neural techniques [3] on human detection employed CNNs or R-CNN [4] object detection architecture, achieves a high-quality externally offered recommendations to attain good performance. In later efforts [5], such YOLOv2, employ single-stage detectors model. While speed is significantly increased by these detectors end-to-end construction in comparison to Faster R-CNN, accuracy is slightly decreased [6]. RetinaNet [7] is a different single-stage detector with detection performance relatively as good as to two-stage methods. A Feature Pyramid Network acts as the backbone on top of a ResNet architecture. Two separate sub-networks classify anchor boxes and modify values in relation to the default anchors. In this study [8], most effective model for identifying the kind of vehicle, each algorithm went through a training dataset of cars and then examined its performance. According to a study, YOLO v3 has advantages in detection speed while maintaining certain MAP i.e. 80.17% MAP (Mean Average Precision) surpasses competing approaches such as Faster R-CNN, SSD where frames per second (FPS) was more than eight times than that of Faster R-CNN. In [9], the author has conducted an experiment to check the viability of utilizing object detection methods to identify safe landing spots in case the UAV suffers an in-flight failure and compared different versions of YOLO model and it shows that YOLOv5 algorithm outperforms YOLOv4 [11] and YOLOv3 in terms of accuracy of detection while maintaining a slightly slower inference speed also a light-weight algorithm to execute worry-free procedure for power-limited UAVs.

This paper [10] inspired the idea of employing HSV color space masking to avoid water bodies and dense vegetation. This technique has no restrictions compared to prior studies. The aerial photos are segmented using color and texture cues to determine acceptable landing places.

#### A. Motivation

The abundance of commercial camera drones served as the motivation for this study. These drones are capable of "return to home" (RTH) and vertical landing. A number of environmental conditions, such as a sudden break in connection between the drone and the controller, instability brought on by a strong wind when the drone is in flight, or lack of power between drone parts, can cause emergencies. In this research, a model is employed for landing camera enabled drones securely without using a target as a reference landing or flying the drones back to their home location, which would be difficult and power intensive. A drone can do a vertical landing using the suggested architecture. According to the survey, YOLOv5 appears to be the most appropriate for real-time processing with a lightweight model for object detection in case of power-limited applications and using grid-based architecture maximizes the area coverage for site selection, as opposed to SLZ candidates, which are obtained as circular regions [15], where valuable portions of an obstacle-free zone may be missed. Our main objective was to offer most commercial drones a stable dynamic landing approach without the use of additional hardware or sensors.

### III. METHODOLOGY

In this paper, an architecture is proposed for autonomous UAV safe landing algorithm. In the event when drone is unable to reach its home location due to low power or emergency, the purpose of our algorithm is to choose a safe landing zone inside populated areas so that it does not cause injury to any of the people within vicinity of drone and also minimal impact to the drone. Specifically, a UAV with a camera mounted on it is being considered. First, the camera is used to detect obstacles using YOLOv5 model. Second, the location of the closest safe landing zone is determined using grid-based architecture. These decisions are then relayed to the PID control block, which generates PID signals for the drone actuators so that they can navigate to the desired location as shown in Fig. 1.

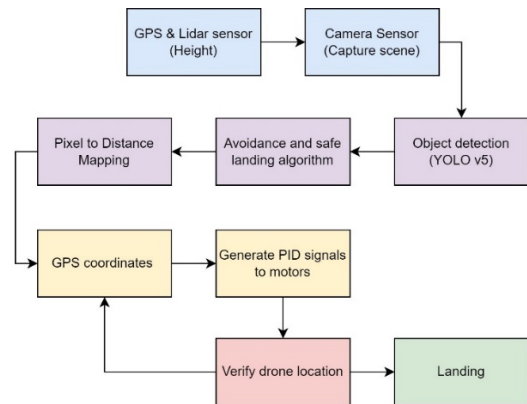


Fig. 1. Block Diagram of Overall Safe Landing Algorithm.

#### A. Camera FOV Distance Calculation

To calculate aerial view coverage of drone using fixed Field of View (FOV) camera is given below.

$$\tan\left(\frac{FOV}{2}\right) \times D \times 2 = N \quad (1)$$

The coverage, denoted by "N," is obtained by using the field of view (FOV) of the camera and the working distance of the drone, "D". Keeping image formats as 1:1 aspect ratio, this results in a 1:1 ratio for the working distance and coverage as shown in Fig. 2(a). The fundamental highlight is the ability to determine the coverage distance based on the height of the drone. Scene 1 in Fig. 2(b) indicates that level-1 operation is being considered at an altitude of 27 meters, and coverage is 27m × 27m. Similarly, at level-2 operation at an altitude of 9m is considered will have a coverage of 9m × 9m shown as scene 2. The aforementioned equation (1) is verified using the practical specifications of a drone camera listed in Table I.

#### B. Object Detection – YOLOv5

Object detecting methods like the single-stage detectors YOLOv5 model are utilized to avoid people, cars, and bikes on metropolitan areas. This model is trained using VisDrone datasets [12] and the Stanford Drone Dataset (SDD) [13]. Basically, YOLO models have three architectural blocks [14].

- YOLOv5 Backbone: It is used to extract image features. YOLO v5 uses CSP (Cross Stage Partial Networks) to obtain useful features from an input image.

- YOLOv5 Neck: It is used to construct feature pyramids. Feature pyramids help scaling models generalize. It helps identify objects in different sizes and scales. Feature pyramids help models perform well on new data. FPN, BiFPN, and PANet use feature pyramids. PANet generates a feature pyramids network to aggregate features and passes it to Head for prediction.
- YOLOv5 Head: Responsible for final detection. It employs anchor boxes to construct class probabilities, objectness scores, and bounding boxes.

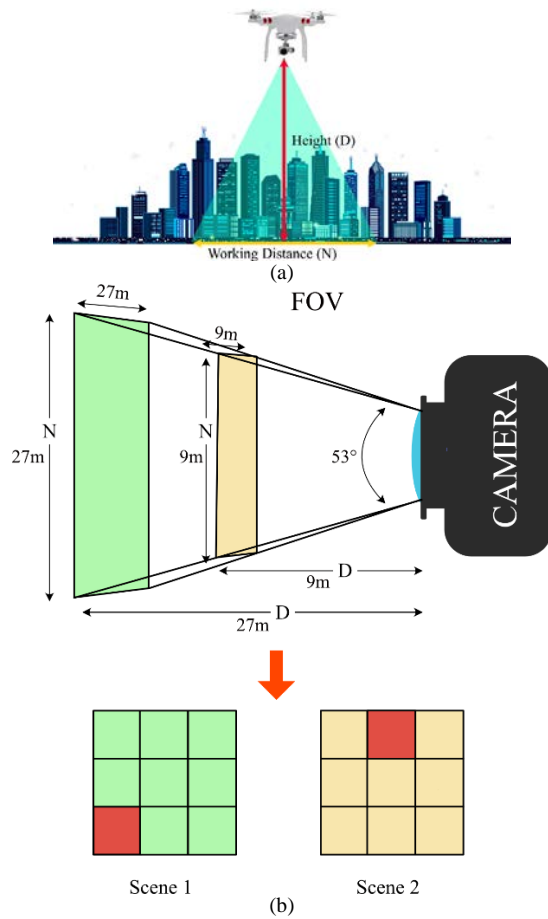


Fig. 2. (a) Drone Representation of Working Distance, (b) Working Distance Calculation.

TABLE I. PARAMETERS FOR DRONE CAMERA

Parameters	Value
Image sensor	1/4"
Image format	1:1 aspect ratio (square)
Image Pixel	1024x1024
FOV	$53^\circ$
Focal length	3.2 mm
working distance	27m
Converge	27m x 27m

### C. Avoidance and Identification of Safe Landing Location

The YOLOv5 output image is then provided to the avoidance system after object detection is done. Here, the image is converted to the HSV format in order to identify YOLO boundary boxes, green vegetation, water bodies such as pool, lake etc. HSV green threshold is then applied as a mask to represent the green vegetation and boundary boxes, while HSV blue represents the water bodies. The HSV range is fine-tuned using threshold of dominant color of image. The area within the boundary boxes is filled to represent an obstacle and then the image is inverted to show obstacles as being black.

The flow diagram as shown in Fig. 3, to determine the safe landing zone, a  $3 \times 3$  grid is applied over the inverted image (scene-1), and at an altitude of 27 meters, each grid has a sufficient area of  $9m \times 9m$  for level-1 operation as show in Fig. 2(b). After a set of safe landing zones (SLZ) are determined and stored in LUT for subsequent use. Therefore, the nearest safe landing area is considered. The geolocation is then determined by doing a pixel to distance mapping and then converting the distance to GPS coordinates. The PID control block receives these GPS coordinates and uses them to provide the necessary signals for the drone actuators to navigate to the specified safe landing site. After descending to 9m from an altitude of 27m, the above set of steps are repeated to locate a sub-zone for the drone to land securely. At altitude 9m with a  $3 \times 3$  grid applied over the scene-2, the drone still has a sufficient space of  $3m \times 3m$  for each grid to land safely. This can be dynamically modified depending on the size or class of the drone.

The below Table II shows that the algorithm is flexible enough to reconsider the decisions while descending to verify if the SLZ selected are truly obstacle free at lower altitudes.

TABLE II. SAFE LANDING DECISION MODES

Mode	No. of SLZ <sub>27m</sub> (At level - 1)	No. of SLZ <sub>9m</sub> (At level - 2)	Response
1	SLZ == 0	Don't care	TRAVERSE to New Site
2	SLZ == 1	SLZ == 0	ASCEND to Level -1; Then TRAVERSE to New Site
3	SLZ > 1	SLZ == 0	ASCEND to Level -1; Then TRAVERSE to Next nearest SLZ
4	SLZ > 0	SLZ > 0	LANDING

### D. Pixel to GPS Coordinates

- Considering the captured scene is of 1:1 aspect ratio selected for computation. Firstly, a distance of 27 meters is mapped onto an image of 1024 pixels by using the map function in Python. Secondly, find the distance from the reference point (center of image) to the desired safe landing zone and calculate the destination GPS coordinate based on the distance to be travelled.

- For distance to GPS coordinates, GPRMC was utilized from the GPS NEMA sentence to obtain coordinates (longitude and latitude information) which decimal degree (DD) format.
- The following formula can be used to convert GPS data in the form of degrees, minutes, and seconds (DMS) to signed decimal degree (DD).

$$\text{GPS coordinate} = \text{degrees} + \frac{\text{minutes}}{60} + \frac{\text{seconds}}{3600} \quad (2)$$

The GPS waypoint distance is calculated using distance module in python using Algorithm 1.

**Algorithm 1** Distance to GPS coordinate Mapping

**Input:** origin: latitude and longitude of current location

D: Distance from origin to destination

sel: To select latitude or longitude

dir: Direction in either West/East or North/South

**Output:**

1: Find destination GPS coordinate lat2 or lon2.

2: Set radius of earth in km, radius = 6371

3: Compute central angle,  $c = D/\text{radius of earth}$

4: **if** (sel is longitude) **then**

5: Compute,

$$dlon = \text{degrees} \left( \cos^{-1} \left[ 1 - \frac{2 \times \tan\left(\frac{c}{2}\right)^2}{\cos(\text{lat}1)^2 \left[ 1 + \tan\left(\frac{c}{2}\right)^2 \right]} \right] \right)$$

6: **if** (dir == North)

7:  $\text{lon}2 = dlon + \text{lon}1$

8: **else if** (dir == South)

9:  $\text{lon}2 = \text{lon}1 - dlon$

10:  $\text{coordinate} = \text{lon}2$

11: **else if** (sel is latitude) **then**

12: Compute,

$$dlat = \text{degrees} \left( \cos^{-1} \left[ 1 - \frac{2 \times \tan\left(\frac{c}{2}\right)^2}{\left[ 1 + \tan\left(\frac{c}{2}\right)^2 \right]} \right] \right)$$

13: **if** (dir == East)

14:  $\text{lat}2 = dlat + \text{lat}1$

15: **else if** (dir == West)

16:  $\text{lat}2 = \text{lat}1 - dlat$

17:  $\text{coordinate} = \text{lat}2$

18: **return** coordinate

This algorithm is inspired using the Haversine equations [16]. This helps to create pre-defined points in grid to send drone to desired location as the scene captured is static.

$$\text{origin} = [\text{lat}1, \text{lon}1] \quad (3)$$

$$\text{destination} = [\text{lat}2, \text{lon}2] \quad (4)$$

$$\delta\text{lat} = \text{radian}(\text{lat}1) - \text{radian}(\text{lat}2) \quad (5)$$

$$\delta\text{lon} = \text{radian}(\text{lon}2) - \text{radian}(\text{lon}1) \quad (6)$$

Using haversine formula,

$$\alpha = \sin^2(\delta\text{lat}/2) + \cos(\text{lat}1) * \cos(\text{lat}2) * \sin^2(\delta\text{lon}/2); \quad (7)$$

$$\phi = 2 * \text{atan} \left( \sqrt{\frac{\alpha}{1 - \alpha}} \right) \quad (8)$$

$$\delta = \text{radius} * \phi \quad (9)$$

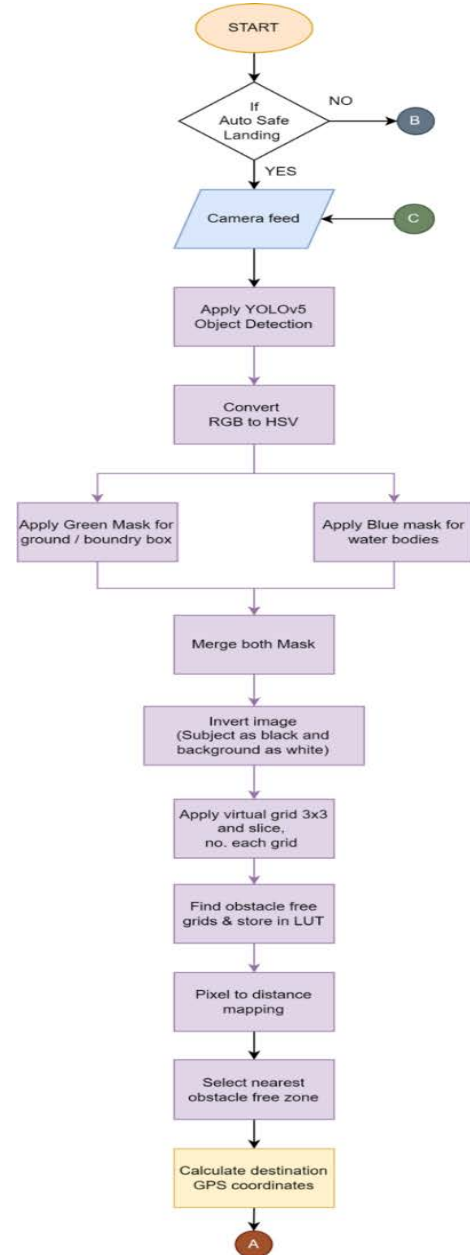


Fig. 3. Functional Flow Diagram of Safe Landing Algorithm.

**E. Generate PID Signals**

The command to send the drone to the desired location is done using only roll (x) and pitch (y) parameters. The current GPS location is assumed and motor control command is initiated upon receiving the safe landing zone GPS coordinates as shown in Fig. 4. In practice, drone angle is calculated from



the gyro rates of the IMU sensor and sent to the PID module to check the angle error to be corrected in the (x,y) region and calculate motor input speed. Then, PWM signals sent to the appropriate drone motors to perform actions such as roll or pitch to cruise to the desired location until the GPS coordinates latitude and longitude matches with destination coordinates. Once the drone reaches the target point, it descends to level-2 height (9m) and repeats the avoidance and location operation for scene 2.

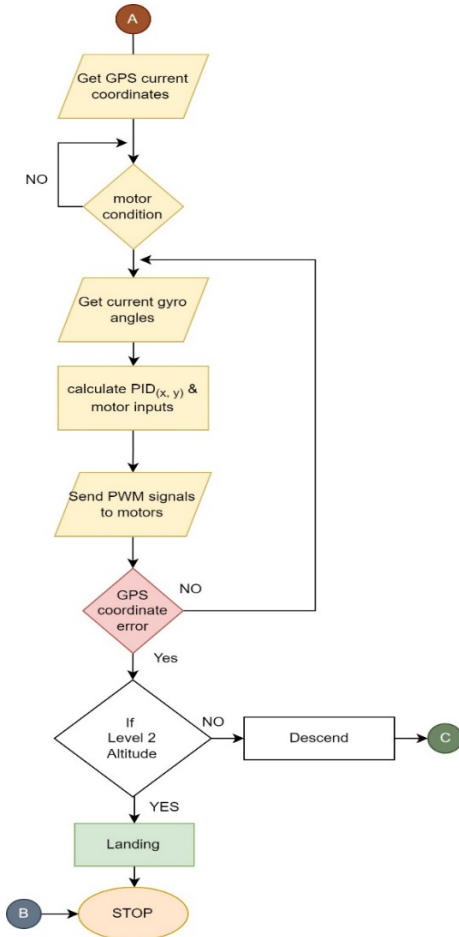


Fig. 4. Functional Flow Diagram of PID Control Block.

#### IV. RESULTS

Utilizing test datasets from VisDrone and SDD pictures as input feed, the proposed safe landing technique is simulated on PC. The following figures illustrate the intermediate results for achieving the safe landing control signals generated for the drone:

##### A. Simulation Results of Avoidance and Safe Landing Location Algorithm

The raw image is fed into YOLOv5 model such that the objects can be identified, and the output of YOLO is shown in Fig. 5 for the objects identified with boundary boxes(green).



Fig. 5. YOLOv5 Output.

In Fig. 6, it shows the output of YOLO is then converted from RGB to HSV image using OpenCV for further image processing. Later, to mask the range of green color a threshold set for HSV image to identify the YOLO's Boundary boxes as well vegetation as shown in Fig. 6(b) where white indicates the masked colors which identifies the boundary boxes and vegetation locations within the captured scene.

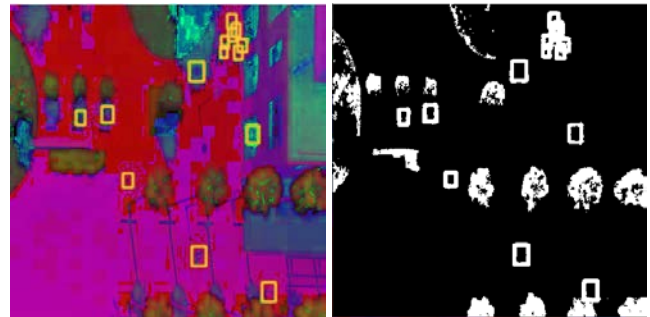


Fig. 6. (a) RGB to HSV Color Space (b) Identify Boundary Boxes.

The next step is that identified contours of the boundary boxes are filled. Then image is inverted, such that to identify the obstacles in black color as shown in Fig. 7.



Fig. 7. (a) Fill in Objects Identified (b) Invert Image.

As shown in Fig. 8 the image is applied with grid such that to indicate each grid is sufficient area for drone to land in that select grid for safe landing area/zone. In this model, 3x3 grid is considered and size of each grid is dependent on the altitude of UAVs. At 27m altitude (level 1 scan), each grid will have as sufficient as 9mx9m grid space. And at 9m altitude (level 2 scan), grid size of 3m x 3m is given for safe landing zone.

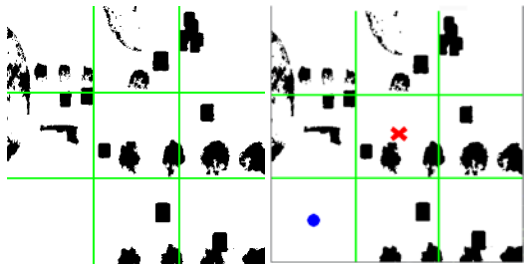


Fig. 8. (a) Apply Grids (b) Safe Landing Zone.

Each grid is split to identify if obstacles are present and is it safe to land. Here all the grids are taken individually and find if any obstacles are there by capturing the black pixels in each grid space and accordingly store all safe landing zones (SLZ) in LUT (Look-up-table). In Fig. 8(b) identifies the nearest SLZ indicated by blue spot which is the shortest distance from the reference point (red cross; resides at the center of image).

### B. Simulation Results of PID Control Signals

The dimensions of image are taken as  $1024 \times 1024$  and assuming that the drone is at 25m height and according to focal length to working distance ratio set to 1:1 ratio the Field of View will be  $25m \times 25m$ , and these values are mapped using map () function. A set of pre-defined values are given for roll/pitch angle (here, roll\_angle =  $12^\circ$ ) which determines the speed of drone, and set until reaches destination waypoint. The below Fig. 9 shows output of each PID for drone motors i.e. roll (x) and pitch (y) plot which indicates that the drone will be navigated to desired safe landing zone location.

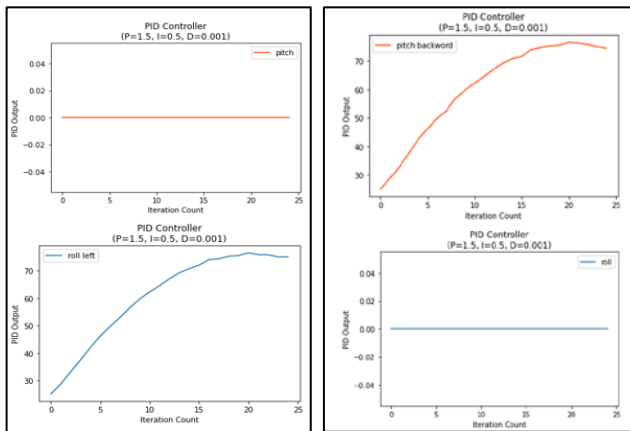


Fig. 9. PID Control Signal – (a) x-axis(roll) (b) y-axis (Pitch).

## V. CONCLUSION

In this paper, the proposed vision-based safe landing algorithm for UAVs intended for urban regions is simulated in Spyder IDE and has been verified using a real-life scene which is taken from SDD and VisDrone test datasets instead of a virtual environment. The custom YOLOv5 is trained for an aerial view of tiny objects such as humans, cars, and bikes and is carried out on a PC with an i9 processor and RTX 3060 GPU specification, which is identified successfully, especially for human detection. The grid-based decision nature of the algorithm for a safe landing will enable maximum coverage of area without missing valuable portions of obstacle-free zone.

By feeding the real-life scenes, the model is verified and successfully works for all possible situations as shown in Table II where the algorithm is flexible enough to make re-decisions if unexpected obstacles occur while descending.

Since the YOLOv5 object detection model allows for the differentiation of objects, this model has a lot of potential for the future, as it can be implemented with a more intelligent system to choose and prioritize where the drone can land without creating any hazards to living beings. In addition, adequate data for tiny object detection allows the decision to be taken from higher altitudes, facilitating the selection of a suitable location for detection and landing.

## REFERENCES

- [1] Symeonidis, Charalampos & Kakaletsis, Efstratios & Mademlis, Ioannis & Nikolaidis, Nikos & Tefas, Anastasios & Pitas, Ioannis. "Vision-based UAV Safe Landing exploiting Lightweight Deep Neural Networks," International Conference on Image and Graphics Processing (ICIGP), pp. 31-19 2021.
- [2] Kakaletsis, E., Tzelepi, M., Kaplanoglou, P. I., Symeonidis, C., Nikolaidis, N., Tefas, A. & Pitas, I. "Semantic map annotation through UAV video analysis using deep learning models in ROS," International Conference on Multimedia Modeling, vol 11296, pp. 328-340, 2019.
- [3] Zhang, S., Benenson, R., Omran, M., Hosang, J., & Schiele, B. "How far are we from solving pedestrian detection?," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1259-1267, 2016.
- [4] Girshick, R., Donahue, J., Darrell, T., & Malik, J. "Rich feature hierarchies for accurate object detection and semantic segmentation," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 580-587, 2014.
- [5] Lan, W., Dang, J., Wang, Y., & Wang, S. "Pedestrian detection based on YOLO network model," IEEE International Conference on Mechatronics and Automation (ICMA), pp. 1547-1551, 2018.
- [6] Ren, S., He, K., Girshick, R. & Sun, J. "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks", Advances in Neural Information Processing Systems (NIPS), vol 28, pp. 91-99, 2015.
- [7] Lin, T.-Y., Goyal, P., Girshick, R., He, K., & Dollar, P. "Focal loss for dense object detection," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 42, no. 2, pp. 318-327, 2017.
- [8] J a. Kim, J. -Y. Sung and S. -h. Park, "Comparison of Faster-RCNN, YOLO, and SSD for Real-Time Vehicle Type Recognition," IEEE International Conference on Consumer Electronics - Asia (ICCE-Asia), pp. 1-4, 2020.
- [9] Nepal, U.; Eslamiat, H., "Comparing YOLOv3, YOLOv4 and YOLOv5 for Autonomous Landing Spot Detection in Faulty UAVs," Sensors 2022, vol. 22, 464, 2022.
- [10] Dehshibi, M.M., Fahimi, M.S., Mashhadi, M. "Vision-Based Site Selection for Emergency Landing of UAVs," Advances in Intelligent Systems and Computing, Springer Cham, vol 361, pp. 397-402, 2015.
- [11] Meena Deshpande, Veena M.B., "License Plate Detection and Recognition using YOLO v4," Webology, vol.18, no.4, 2021.
- [12] P. Zhu et al., "Detection and Tracking Meet Drones Challenge," IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021.
- [13] A. Robicquet, A. Sadeghian, A. Alahi, S. Savarese, "Learning Social Etiquette: Human Trajectory Prediction In Crowded Scenes" in European Conference on Computer Vision (ECCV), vol 9912, pp. 549-565, 2016.
- [14] YOLOv5 Glenn Jocher, Ayush Chaurasia, Alex Stoken, Jirka Borovec, NanoCode012, Yonghye Kwon, TaoXie, Kalen Michael, Jiacong Fang, imyhxy, Lorna, Colin Wong, Zeng Yifu, Abhiram V, Diego Montes, Zhiqiang Wang, Cristi Fati, Jebastin Nadar, Laughing, ... xylieong. (2022). ultralytics/yolov5: v6.2 - YOLOv5 Classification Models, Apple M1, Reproducibility, ClearML and Deci.ai integrations (v6.2). Zenodo. <https://doi.org/10.5281/zenodo.7002879>.

- [15] J. González-Trejo, D. Mercado-Ravell, I. Becerra and R. Murrieta-Cid, "On the Visual-Based Safe Landing of UAVs in Populated Areas: A Crucial Aspect for Urban Deployment," in *IEEE Robotics and Automation Letters*, vol. 6, no. 4, pp. 7901-7908, Oct. 2021.
- [16] Gong, Yikai & Deng, Fengmin & Sinnott, Richard "Identification of (near) Real-time Traffic Congestion in the Cities of Australia through Twitter", *ACM First International Workshop*, pp. 7-12, 2015.
- [17] Anne Goodchild & Jordan Toy, "Delivery by drone: An evaluation of unmanned aerial vehicle technology in reducing CO2 emissions in the delivery service industry," *Transportation Research Part D: Transport and Environment*, vol 61, Part A, pp. 58-67, 2018.
- [18] Arfaoui, Aymen. "Unmanned Aerial Vehicle: Review of Onboard Sensors, Application Fields, Open Problems and Research Issues," *International Journal of Image Processing*, vol 11, pp. 12-24, 2017.

# Gaussian Projection Deep Extreme Clustering and Chebyshev Reflective Correlation based Outlier Detection

S. Rajalakshmi<sup>1</sup>

Research Scholar

Department of Computer Science  
Periyar University, Salem, Tamilnadu, India

Dr. P. Madhubala<sup>2</sup>

Research Supervisor

Department of Computer Science  
Periyar University, Salem, Tamilnadu, India

**Abstract**—Outlier detection or simply the task of point detection that are noticeably distinct and different from data sample is a predominant issue in deep learning. When a framework is constructed, these distinctive points can later lead to model training and compromise accurate predictions. Owing to this reason, it is paramount to recognize and eliminate them before constructing any supervised model and this is frequently the initial step when dealing with a deep learning issue. Over the recent few years, different numbers of outlier detector algorithms have been designed that ensure satisfactory results. However, their main disadvantages remain in the time and space complexity and unsupervised nature. In this work, a clustering-based outlier detection called, Random Projection Deep Extreme Learning-based Chebyshev Reflective Correlation (RPDEL-CRC) is proposed. First, Gaussian Random Projection-based Deep Extreme Learning-based Clustering model is designed. Here, by applying Gaussian Random Projection function to the Deep Extreme Learning obtains the relevant and robust clusters corresponding to the data points in a significant manner. Next, with the robust clusters, outlier detection time is said to be reduced to a greater extent. In addition, a novel Chebyshev Temporal and Reflective Correlation-based Outlier Detection model is proposed to detect outliers therefore achieving high outlier detection accuracy. The proposed approach is validated with the NIFTY-50 stock market dataset. The performance of the RPDEL-CRC method is evaluated by applying it to NIFTY-50 Stock Market dataset. Finally, we compare the results of the RPDEL-CRC method to the state-of-the-art outlier detection methods using outlier detection time, accuracy, error rate and false positive rate evaluation metrics.

**Keywords**—Outlier detection; clustering; Gaussian random projection; deep extreme learning; Chebyshev distance; temporal; reflective correlation

## I. INTRODUCTION

Outliers are nothing but considered as data points or observations that plunge extraneous of an anticipated distribution or pattern and hence considered as the most-hottest topics as far as data mining is concerned. For example, if we were to perform data approximation with a Binomial distribution, then the outliers are the findings that do not emerge to go along with the pattern of a Binomial distribution. It discover anomalous data objects and are said to be of high use in several applications like detecting network intrusion, detecting fraudulent activities concerning credit card

management, outlier detection in stock market to mention few. In the area of outlier detection, the ground truth is found to be seldom missing and hence machine learning techniques are extensively utilized in outlier detection research.

Most of the prevailing research works concentrates on outlier detection for categorical or numerical attribute data. A fuzzy rough set (FRSs) was proposed in [1] to detect outlier in mixed attribute data based on fuzzy rough granules. Initially, the granule outlier degree (GOD) was designed with the objective of characterizing the outlier degree of fuzzy rough granules via fuzzy approximation accuracy.

Followed by which, the outlier factor on the basis of fuzzy rough granules was designed by integrating GOD and respective weights to measure outlier degree of objects using fuzzy rough granules-based outlier detection (FRGOD) algorithm. With this both precision and recall were said to be improved. Despite improvement observed in terms of precision and recall, the time and space complexity were relatively high. To address on this aspect, Gaussian Random Projection-based Deep Extreme Learning-based Clustering model is first designed and then the outliers are detected. With this process, the time and space complexity involved in outlier detection will be reduced to a greater extent.

Iterative ensemble method with distance-based data filtering was proposed in [2] based on an iterative approach with the purpose of detecting outliers present in unlabeled data. The ensemble method was utilized in clustering the unlabeled data. Then, with the clustered data potential outliers were filtered in an iterative manner employing cluster membership threshold. This was performed in an iterative manner until Dunn index score for clustering was said to be maximized.

On the other hand, the distance-based data filtering eliminated the prospective outlier clusters from post-clustered data on the basis of the distance threshold utilizing the Euclidean distance measure from majority cluster as filtering factor. With this the improvement were found to be observed in terms of both precision and f-score value. Despite improvement observed in terms of both precision and f-score, by detecting possible outlier clusters based on weighted method, the false positive rate can be reduced to a greater extent. With this objective, Chebyshev Temporal and Reflective Correlation-based Outlier Detection model is

designed so that using Chebyshev distance based temporal factor obtains highly correlated data points, therefore reducing the false positive rate to a greater extent.

#### A. Objective and Contributions

The main objective of this research is to propose a novel cluster-based outlier detection method that performs clustering process and outlier detection separately in a significant manner. This clustering-based outlier detection method addresses the limitations of the earlier outlier detection methods by its multi-factor i.e., deep clustering and correlative outlier detection model. Further, the contributions of this paper include the following.

- To propose a novel Gaussian Random Projection-based Deep Extreme Learning-based Clustering algorithm to minimize a composite objective function, i.e., outlier detection time along with the improvement in error rate. The model minimizes the outlier detection time and reduces error rate during outlier detection via two different functions, Gaussian Random Projection and square gradient function.
- To design a new Chebyshev Temporal and Reflective Correlation-based Outlier Detection algorithm based on Chebyshev Temporal function and Reflective Correlative function that ensures accurate outlier detection.
- The proposed Random Projection Deep Extreme Learning-based Chebyshev Reflective Correlation (RPDEL-CRC) method has provided improved results for outlier detection time, accuracy, error and false positive rate as performance evaluation measures.

#### B. Organization of the Paper

The rest of the paper is organized as: The discussion about the obtainable modern outlier detection techniques is presented in Section II. In Section III, the proposed Random Projection Deep Extreme Learning-based Chebyshev Reflective Correlation (RPDEL-CRC) method has been discussed. In Section IV, Chebyshev temporal and reflective correlation-based outlier detection model is discussed. The discussion about the experimental setup and comparative analysis with an elaborate discussion is described in Section V and finally, the conclusions are presented in Section VI.

## II. RELATED WORKS

The issue of outlier detection consists of detecting and eliminating malicious inferences from data. This problem is found to take place in several applications. However, outliers are frequently equipped by data stream that in turn influence the accuracy of data-based predictions. Hence, there arises an acute requirement to identify the outliers so as to enhance the data reliability.

A novel method to identify trajectory outlier group from large trajectory database using different types of algorithms was proposed in [3]. First, algorithms based on data mining were designed to identify the correlations between trajectory data and identify abnormal trajectories. Second, machine learning algorithms were applied to identify the group of

trajectory outliers. Finally, convolution deep neural network were used to learn distinct different features to determine group of trajectory outliers, therefore enhancing runtime and accuracy performance.

Conventional outlier detection method however does not take into consideration the subset occurrence frequency and hence, the outliers being detected do not fit the definition of outliers. To address on this aspect, a two-phase minimal weighted rare pattern mining-based outlier detection method, called MWRPM-Outlier [4] was proposed to efficiently detect outliers based on the weight data stream.

A novel methodology to identify conjunct unusual human behaviors from large pedestrian data in smart cities was proposed in [5]. First, data mining was used followed by which convolution deep neural networks was explored that in turn identified distinct features to determine collective abnormal human behavior. With this both runtime and accuracy were said to be improved.

Despite several outlier detection algorithms are said to exist for scenarios necessitating numerical data, only a few prevailing methods can control categorical data. Moreover, the methods outlined for categorical data severely endure from two issues, low detection precision and high time complexity. Two novel outlier detection mechanisms for categorical data sets were proposed in [6]. First an entropy based method using Outlier Detection Tree (ODT) was designed followed by which second simple if-then rules were utilized for outlier detection. With these two integrated mechanisms both precision and computational complexity were improved to a greater extent.

Outlier detection has received paramount significance in the domain of data mining owing to the requirement to detect unusual events in different types of applications, to name a few being, fraud detection, intrusion detection and so on. Different types of outlier detection algorithms have been proposed in the recent past for utilization on static data sets employing a finite number of samples.

Probabilistic deep autoencoder was proposed in [7] with the objective of reconstructing measurements of power system that in turn can be employed in outlier detection. First, nonparametric distribution estimation method was utilized for obtaining information pertaining to uncertainty. Second, confidence intervals were acquired from estimated distribution and were further utilized as input. Finally, based on the multilayer encoding and decoding processes, the measurement intervals were reconstructed, with which outlier detection were made in an accurate manner.

Outlier detection methods employing machine learning are said to be receiving greater attention in the past few years in several domains. But, an ensemble of such outlier detection methods could improve the overall detection performance. An algorithm called, Average Selection and Ensemble of Candidates for Outlier Detection (DASEC-OD) was proposed in [8] for high dimensional data. A review of unsupervised outlier detection methods focusing on multi-dimensional data was investigated in [9].

With the exponential requirement in analyzing high speed data streams, the job of outlier detection becomes more



demanding as the conventional outlier detection method can no longer presume all data for processing. In [10], a Memory-efficient incremental Local Outlier (MiLOF) was proposed for large data streams, therefore ensuring accuracy to a greater extent.

Nowadays, there prevail very huge types of outlier detector methods that bestow satisfactory results. But their major disadvantage remains in their unsupervised characteristic in conjunction with the hyper parameters that has to be appropriately assigned for acquiring better performance.

An improved content-based outlier detection method was proposed in [11]. In [12], a novel supervised outlier estimator was designed. This was performed by pipelining an outlier detector in such a manner that the targets involved in the outlier detector were obtained in an optimal manner. However, these methods did not perform in a satisfactory manner in case of utilization of the complex datasets and hence suffer from noise introduced by outliers, specifically when the ratio of outlier was found to be high. To address this aspect, a framework called, Transformation Invariant AutoEncoder (TIAE) was proposed in [13] that in turn attained not only stability but also ensured high performance on outlier detection. A comprehensive review of outlier detection techniques were investigated in [14].

In several practical classification issues, a portion of outliers are said to exist in datasets that in turn would have heavy influence on the constructed model performance. A group method of data handing (GMDH) using neural network in outlier detection was proposed in [15]. A novel robust outlier detection method (RiLOF) based on Median of Nearest Neighborhood Absolute Deviation (MoNNAD) was designed in [16] that employed median of local absolute deviation of the samples to attain high detection performance.

Monitoring data including the significant information of monitored object forms the fundamentals for data mining and analysis. However, the data being monitored suffers from outlier pollution therefore causing negative influence on corresponding data processing. To address on this aspect, an outlier detection method based on stacked autoencoder (SAE) was proposed in [17]. The proposed SAE had the significant potentiality of feature extraction and heavily maintained the indigenous information of data to a greater extent.

Accuracy and time involved in outlier detection was not focused. To address this aspect, a Neighbor Entropy Local Outlier Factor was presented in [18] that with the aid of self organizing feature map not only improved accuracy but also reduced the execution time to a greater extent. Moreover, semantic information was focused on [19] for outlier detection employing meta path based outlier detection. Outlier detection based on the multivariable panel data was designed in [20] via correlation coefficient that in turn indicated high accuracy detection ability.

Motivated by the above mentioned techniques in this work, a novel cluster-based outlier detection method called, Random Projection Deep Extreme Learning-based Chebyshev Reflective Correlation is proposed (RPDEL-CRC). The

elaborate description of RPDEL-CRC method is presented in the following sections.

### III. RANDOM PROJECTION DEEP EXTREME LEARNING-BASED CHEBYSHEV REFLECTIVE CORRELATION (RPDEL-CRC)

The proposed Random Projection Deep Extreme Learning-based Chebyshev Reflective Correlation method concentrates on the detection of outliers based on clustering. Methods designed based on cluster detect the outliers by placing data objects into distinct clusters. Here, the data objects in a data set are initially clustered. To design cluster-based outlier detection, the RPDEL-CRC method is split into two parts. Fig. 1 shows the block diagram of RPDEL-CRC method.

As illustrated in the figure below, the first part models robust cluster by means of Gaussian Random Projection-based Deep Extreme Learning. Here, the clustering based outlier detection initiates the outlier detection process by clustering the given input dataset, Nifty 50 Stock Market Data (2000 – 2021). Hence, to be more specific outliers are considered as data points within deviating clusters or the data points that deviate to the formed clusters. The second part uses the Chebyshev Temporal and Reflective Correlation-based Outlier Detection algorithm to detect outlier with minimum falsification. In this section, we first explain all prerequisites of the proposed method with a system model, and then finally we describe the proposed method.

#### A. System Model

Let  $P \in R^{(m*n)}$  represent a matrix with 'm' rows and 'n' columns of real numbers  $P_{ij} \in R$ . The matrix 'P' denotes a dataset 'DS' that includes the data for outlier analysis. The 'n' columns are called features and on the other hand, the 'm' are referred to as data points. Then, vector  $[DP]_i \in R^n$  refers to the data point, which is a row in 'P'. The matrix 'P' then consists of 'm' data points  $DP = \{ [DP]_1, DP_2, \dots, [DP]_n \}$ . Then, with the aid of the outlier detection algorithm the outliers present in the dataset 'DS' are detected. Finally, the overall feature space represents the vector space defined by the given features that in turn estimates the characteristics of the examined occurrence or event. Inliers are detected in subsets of the overall feature space and referred to as normal regions or normal data points. To be more specific, inliers are considered to as the data points in the normal regions. On the other hand, an outlier is a data point that does not belong in the normal region.

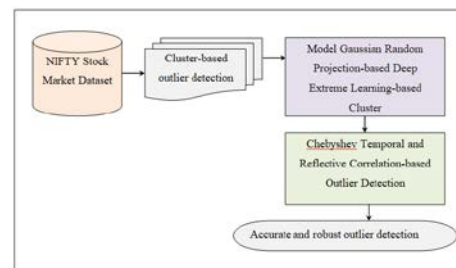


Fig. 1. Block Diagram of Random Projection Deep Extreme Learning-based Chebyshev Reflective Correlation Method.



### B. Case Analysis of Outlier Detection Accuracy

To detect outliers based on the cluster in a given dataset, the data has to be initially clustered. In this paper, Gaussian Random Projection-based Deep Extreme Learning model is first employed for clustering. The objective behind the design of Gaussian Random Projection-based Deep Extreme Learning model remains in training feed forward network from a raw training data set with ‘N’ samples, ‘{P,Q}={P<sub>i</sub>,Q<sub>i</sub> }<sub>(i=1,2,3, … ,N)</sub>’, with ‘P<sub>i</sub> ∈ R<sup>d</sup>’ and ‘Q<sub>i</sub>’ represents ‘M-dimensional’ binary vector where one entry denotes ‘1’ representing the cluster that ‘P<sub>i</sub>’ belongs to. Fig. 2 shows the block diagram of Gaussian Random Projection-based Deep Extreme Learning-based Clustering model.

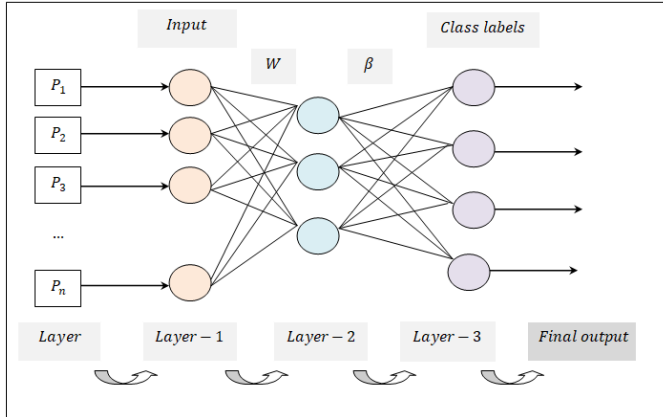


Fig. 2. Block Diagram of Gaussian Random Projection-based Deep Extreme Learning-based Clustering Model.

As shown in the above figure, the training process of GRP-DEL includes two steps. In (1), the hidden layer with ‘K’ nodes employing distinct numbers of neurons are constructed. Next, for the ‘i-th’ hidden layer node, a ‘d-dimensional’ vector ‘x<sub>j</sub>’ and a metric ‘y<sub>j</sub>’ are generated in an arbitrary manner. Then, for each input vector ‘P<sub>i</sub>’, the pertinent output on the ‘i-th’ hidden layer node is obtained by utilizing Sigmoid activation function. This is mathematically stated as given below.

$$g(P_i, x_j, y_j) = \frac{1}{1 + \exp[-(x_j^T * P_i + y_j)]} \quad (1)$$

Then, using the resultant value of the above Sigmoid activation function, the hidden layer outputs the matrix as given below.

$$H = \begin{bmatrix} g(P_1, x_1, y_1) & \dots & g(P_1, x_K, y_K) \\ \dots & \dots & \dots \\ g(P_N, x_1, y_1) & \dots & g(P_N, x_K, y_K) \end{bmatrix}_{N \times K} \quad (2)$$

In (2), an ‘M-dimensional’ binary vector ‘α<sub>j</sub>’ represents the output weight that associates the ‘j-th’ hidden layer with the resultant output node. Here, a random projection is applied that states that if points associating the ‘j-th’ hidden layer in a vector space are of sufficiently high dimension, then the ‘j-th’ hidden layer may be projected into a lower-dimensional space in such a manner that it preserves the distances between points (therefore minimizing dimensionality). With original input vector being ‘P<sub>i</sub> (K\*M)’, using a random ‘K\*d’ matrix dimensional matrix ‘R’, then the projection of data on to a

lower dimensional subspace is mathematically formulated as given below.

$$P_{K \times M} = R_{K \times d} P_{d \times M} \quad (3)$$

Then, with the above lower dimensional subspace random projection dimensionality of set of points and output matrix ‘Q’ is mathematically stated as given below.

$$H \cdot \alpha = Q \quad (4)$$

$$\alpha = \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \dots \\ \alpha_K \end{bmatrix}_{K \times M}; Q = \begin{bmatrix} Q_1 \\ Q_2 \\ \dots \\ Q_N \end{bmatrix}_{N \times M} \quad (5)$$

Next, with the resultant matrices ‘H’ and ‘Q’, the objective of GRP-DEL model remains in solving the output weights ‘α’ by reducing the losses of prediction errors, leading to the following equation.

$$\alpha_i(n) = \alpha_i(n-1) - \beta_i(n) \frac{MA_i(n)}{\sqrt{G_i(n)}} \quad (6)$$

From the above (6), ‘[[MA]]<sub>i</sub>(n)’ symbolizes the moving average of feature or attribute ‘i’ at iteration ‘n’, with square gradient denoted by ‘G<sub>i</sub>(n)’ and learning rate ‘β<sub>i</sub>(n)’ respectively.

$$MA_i(n) = \gamma_n MA_i(n-1) + (1 - \gamma_n) \quad (7)$$

$$G_i(n) = \theta_n G_i(n-1) + (1 - \theta_n) \quad (8)$$

$$\beta_i(n) = \beta_i(n-1) \frac{\sqrt{(1-\theta_n)^n}}{(1-\gamma_n)^n} \quad (9)$$

From the above (7), (8) and (9) the factors ‘γ<sub>n</sub>’ and ‘θ<sub>n</sub>’ are utilized in fine tuning the decay rates of moving averages close to one (i.e., ‘γ<sub>n</sub>=0.85’ and ‘θ<sub>n</sub>=0.9’). The pseudo code representation of Gaussian Random Projection-based Deep Extreme Learning-based Clustering is given below.

Algorithm 1: Gaussian Random Projection-based Deep Extreme Learning-based Clustering

---

**Input:** Dataset ‘DS’, data points ‘DP = {DP<sub>1</sub>, DP<sub>2</sub>, …, DP<sub>n</sub>}’  
**Output:** obtain cluster ‘Q<sub>i</sub>’ corresponding to ‘P<sub>i</sub>’ in computationally efficient and precise manner

---

- 1: **Initialize** ‘m’ rows and ‘n’ columns
  - 2: **Begin**
  - 3: **For** each Dataset ‘DS’ with data points ‘DP’ and input vector ‘P<sub>i</sub>’
  - 4: Obtain pertinent output on the ‘i-th’ hidden layer employing Sigmoid activation function as in (1)
  - 5: Obtain output matrix via hidden layer as in (2)
  - 6: Evaluate Gaussian Random Projection as in (3)
  - 7: Estimate hidden layer output and calculate the output matrix as in (4) and (5)
  - 8: **Repeat** (training of neural networks)
  - 9: Solve output weights by minimizing prediction loss error as in (6)
  - 10: Treat each row of ‘Q’ as a point and cluster them into ‘K’ clusters
  - 11: Estimate learning rates for cluster parameter as in (7), (8) and (9)
  - 12: **Until** (first-order gradients for neural networks is arrived at)
  - 13: **Return** ‘Q’
  - 14: **End for**
  - 15: **End**
-

As given in the above Gaussian Random Projection-based Deep Extreme Learning-based Clustering algorithm, with the objective of reducing the outlier detection time along with the improvement in precision, two different functions are employed. First, by employing Gaussian Random Projection the dimensionality of data is said to be reduced by projecting original input space (i.e., the raw data) with the aid of a sparse random matrix. Second, by estimating the learning rate by means of square gradient minimizes the error involved during the process of clustering to a greater extent. As a result, with these two function, clusters are formed both in a computationally efficient and precise manner.

#### IV. CHEBYSHEV TEMPORAL AND REFLECTIVE CORRELATION-BASED OUTLIER DETECTION MODEL

Outlier detection remains to be one of the primary step in data mining tasks. The motive behind the outlier detection strategy here is to identify the features or parameters that are counterfeit from several other features. Different types of outlier detection models are said to exist. In order to determine the perpetual temporal outliers, we obtain outliers based on distance measures by analyzing temporal values of the objects employing Chebyshev Temporal and Reflective Correlation-based Outlier Detection model. Fig. 3 shows the block diagram of Chebyshev Temporal and Reflective Correlation-based Outlier Detection model.

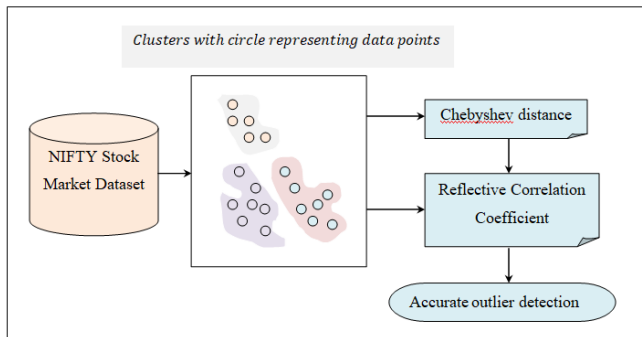


Fig. 3. Block Diagram of Chebyshev Temporal and Reflective Correlation-based Outlier Detection Model.

As shown in the above figure, with the obtained clusters for the given dataset ‘DS’, in a ‘d-dimensional’ vector, with data point denoted by ‘DP={DP[1],DP[2],...DP[d]}’ at time instance ‘T’, distance between two points ‘[[DP]]\_i’ and ‘DP\_j’ employing Chebyshev distance is mathematically, expressed as given below.

$$Dis(DP_i, DP_j) = Max(|DP_i - DP_j|) \quad (10)$$

From the above (10), by employing the Chebyshev distance measure ‘Dis ([[DP]]\_i, DP\_j)’, the greatest difference between two vectors (i.e., the data points) along any coordinate dimension (i.e., the cluster) is evaluated based on the maximum ‘Max(|[[DP]]\_i- [[DP]]\_j|)’ distance along one axis. To be more specific, based on the principle of chessboard distance as the minimum number of moves required by a king to go from one square to another is utilized; by means of Chebyshev distance, the overall outlier detection accuracy is said to be improved. Next, with the assumption of ‘m’ clusters

‘[[CI]]\_1, [[CI]]\_2,..., [[CI]]\_m’ the centroid data point ‘CP’ is then measured as given below.

$$Cl_i CP [i] = \frac{\sum_{P \in Cl_i} DP[i]}{|Cl_i|} \quad (11)$$

With the above determination of the centroid (11), with the assumption that in a cluster ‘CI’ most of the normal data points are hardly encircling the centroid data point of cluster ‘CI’, the abnormal data points or the outlier are those generally farther from the centroid data point. Then the updated weight of a centroid data point is mathematically stated as given below.

$$W(DP) = \frac{Dis(DP, Cl_i CP [i])}{\sum_{S \in Neigh(DP)} Dis(S, Cl_i CP [i])} \quad (12)$$

From the above (12), the weight of data point ‘DP’ in cluster ‘[[CI]]\_i’ is estimated based on the neighbors of ‘Neigh(DP)’ in ‘[[CI]]\_i’. Finally, to reflect the robustness and direction of linear correlation between two data points and minimizing the false positive cases, Reflective Correlation Coefficient is applied. RCC function is employed to evaluate the amount of dependency between two distributions of normalized scores ‘G\_i^Norm (DP), G\_j^Norm (DP)’ and is mathematically stated as given below.

$$RCC(DP_i, DP_j) = \frac{W(DP_i)W(DP_j)}{\sqrt{(\sum DP_i)^2 (DP_j)^2}} \quad (13)$$

From the above (13), reflective correlation coefficient ‘RCC’ is obtained based on the weighted data points ‘W([[DP]]\_i)’ and ‘W([[DP]]\_j)’ respectively. The final form of the objective function for minimizing the false positive cases of the ‘j-th detector’ is mathematically stated as given below.

$$Res_j = [G_j^{Norm}(DP)] - [G_j^{Norm}(DP_o)] \quad (14)$$

From the above (14) ‘G\_j^Norm’ forms the normalized score function of the ‘j-th detector’, ‘DP’ denoting the data points with contaminated dataset and ‘[[DP]]\_o’ denoting the outliers. The pseudo code representation of Chebyshev Temporal and Reflective Correlation-based Outlier Detection is given.

#### Algorithm 2 Chebyshev Temporal and Reflective Correlation-based Outlier Detection

<b>Input:</b> Dataset ‘DS’, data points ‘DP = {DP <sub>1</sub> , DP <sub>2</sub> , ..., DP <sub>n</sub> }’
<b>Output:</b> Accurate Outlier Detection
1: <b>Initialize</b> time instance ‘T’
2: <b>Begin</b>
3: <b>For</b> each Dataset ‘DS’ with data points ‘DP’ and cluster ‘Q <sub>i</sub> ’
4: Evaluate distance between data points ‘DP <sub>i</sub> ’ and ‘DP <sub>j</sub> ’ as in (10)
5: <b>For</b> each cluster ‘Q <sub>i</sub> ’
6: Evaluate centroid data point as in (11)
7: Evaluate weight of data point as in (12)
8: Estimate Reflective Correlation Coefficient as in (13)
9: Obtain final form of the objective function of the ‘j – th detector’ as in (14)
10: <b>Return</b> (outliers ‘DP <sub>o</sub> ’)
11: <b>End for</b>
12: <b>End for</b>
13: <b>End</b>

As given in the above Chebyshev Temporal and Reflective Correlation-based Outlier Detection algorithm, with the objective of improving the outlier detection accuracy with minimum falsification, two different functions are employed. First with the obtained clusters based on the data points, Chebyshev distance function is applied to estimate the difference between two data points along any cluster. Based on the minimum number of positioning between clusters, according to time, results are obtained, therefore ensuring outlier detection accuracy. Second by employing the Reflective Correlation Coefficient function dependency between two distributions or data points are obtained therefore reducing the false positive rate to a greater extent. Finally, the outliers are obtained.

### V. EXPERIMENTAL SETUP

In this section, experimental analysis of the Random Projection Deep Extreme Learning-based Chebyshev Reflective Correlation (RPDEL-CRC) method for outlier detection in data mining is presented. In this section, the performance of the proposed RPDEL-CRC is compared with the state-of-the-art methods, fuzzy rough granules-based outlier detection (FRGOD) [1] and Iterative ensemble method with distance-based data filtering [2] using NIFT-50 Stock Market Dataset (<https://www.kaggle.com/datasets/rohanrao/nifty50-stock-market-data>). Simulations are performed in R Programming language. Fair comparison between proposed RPDEL-CRC method and existing fuzzy rough granules-based outlier detection (FRGOD) [1] and Iterative ensemble method with distance-based data filtering [2] are made for evaluating different parameters like, outlier detection time, outlier detection accuracy, false positive rate and error for different iterations.

#### A. Case Analysis of Outlier Detection Time

The first metric significant for cluster based outlier detection is the time consumed in detecting the outlier. To be more specific, outlier detection time refers to the time consumed in detecting the outliers. Lower the outlier detection time more efficient the method is said to be because earlier the time consumed in detecting the outlier is more efficient the method is. The outlier detection time is mathematically stated as given below.

$$OD_{time} = \sum_{i=1}^n Samples_i * Time [Res_j] \tag{15}$$

From the above (15), the outlier detection time ‘[[OD]]\_time’ is measured based on the samples involved in the simulation process ‘[[Samples]]\_i’ and the time consumed in detecting the outliers ‘Time [[Res]]\_j’. It is measured in terms of milliseconds (ms). Table I given below shows the results of outlier detection time observed for three different methods, RPDEL-CRC, FRGOD [1] and Iterative ensemble method with distance-based data filtering [2].

Fig. 4 illustrated above shows the outlier detection time with respect to 50000 different numbers of samples obtained at different intervals from different companies stock values between years 2007 and 2021. These curves are plotted with increasing cardinality of training samples ranging between 5000 and 50000. With the increasing cardinality, the number of

samples involved in analysis for outlier detection increases and therefore an increase in the outlier detection time is observed. However, simulations with 5000 samples observed ‘250ms’ for detecting outliers with respect to single stock sample using RPDEL-CRC, ‘350ms’ for detecting outliers with respect to single stock sample using [1] and ‘450ms’ for detecting outliers with respect to single stock sample using [2]. From this analysis it is inferred that the outlier detection time using RPDEL-CRC is comparatively lesser than [1] and [2]. The reason behind is the incorporation of Gaussian Random Projection-based Deep Extreme Learning-based Clustering model. By applying this model, dimensionality of data or data points are said to be reduced using Gaussian Random Projection based on projecting original input space (i.e., the raw data) with the aid of a sparse random matrix. With this, data points considered to be outliers are obtained that in turn assist in detecting the outliers altogether. Therefore, the outlier detection time using RPDEL-CRC method is found to be reduced by 20% compared to [1] and 37% compared to [2].

TABLE I. TABULATION FOR OUTLIER DETECTION TIME

Samples	Outlier detection time (ms)		
	RPDEL-CRC	FRGOD	Iterative ensemble method with distance-based data filtering
5000	250	350	450
10000	295	395	555
15000	355	435	635
20000	410	485	680
25000	435	535	745
30000	485	625	795
35000	525	685	835
40000	595	745	890
45000	685	800	920
50000	735	835	955

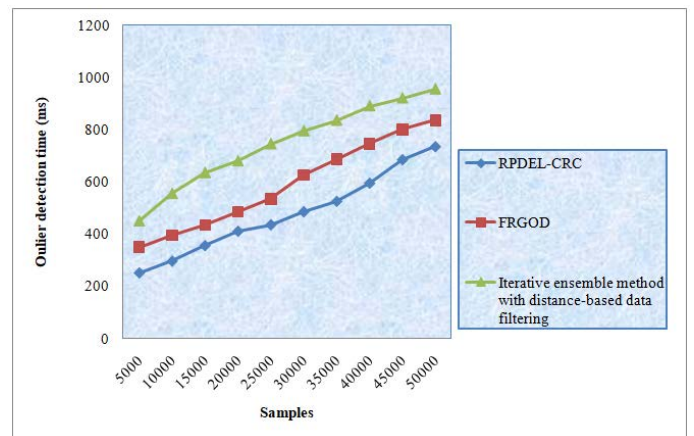


Fig. 4. Graphical Representation of Outlier Detection Time.

#### B. Case Analysis of Outlier Detection Accuracy

The second parameter of significance for cluster based outlier detection is the accuracy rate. In other words, the outlier detection accuracy is measured based on percentage ratio

between the samples involved in simulation process ‘ $\llbracket$  Samples $\rrbracket_i$ ’ and the actual samples accurately detected with outliers ‘ $\llbracket$  Samples $\rrbracket_{ODC}$ ’. This is mathematically stated as given below.

$$OD_{acc} = \sum_{i=1}^n \frac{Samples_{ODC}}{Samples_i} * 100 \quad (16)$$

Table II given shows the results of outlier detection accuracy observed for three different methods, RPDEL-CRC, FRGOD [1] and Iterative ensemble method with distance-based data filtering [2].

Fig. 5 illustrates the outlier detection accuracy for 50000 different stock samples obtained from the NIFTY-50 stock dataset at different time instances. From the figure it is inferred that the outlier detection accuracy is found to be inversely proportional to the stock samples involved in the simulation process. In other words, increasing the stock samples for detecting the outlier causes an increase in the overall data points involved in the process and this in turn minimizes the outlier detection accuracy. However, sample simulations performed with 5000 samples 4845 samples were accurately detected with outliers as it is using RPDEL-CRC, 4755 samples using [1] and 4695 samples using [2]. With this the overall accuracy using the three methods were found to be 96.90%, 95.1% and 93.9% respectively. The overall accuracy was found to be improved using RPDEL-CRC upon comparison with [1] and [2]. The reason behind the outlier detection accuracy improvement was owing to the application of Chebyshev distance function. By applying this distance function, the difference between two data points along any cluster was first evaluated. Then, on the basis of the minimum number of positioning between clusters, according to time, results were obtained, i.e., outliers were detected, therefore ensuring outlier detection accuracy. This in turn improved the outlier detection accuracy using RPDEL-CRC method by 3% compared to [1] and 7% compared to [2].

### C. Case Analysis of False Positive Rate

False positive rate is measured as the ratio between the numbers of negative events (i.e., negative outliers) wrongly categorized as positive (i.e., outliers) and the total number of actual negative events (i.e., actual outliers). This is mathematically stated as given below.

$$FPR = \frac{FP}{FP+TN} \quad (17)$$

From the above (17), the false positive rate ‘FPR’ is measured based on the false positive samples ‘FP’ (i.e., actually the data are not outliers) and the true negative samples ‘TN’ (i.e., outliers detected as outliers) respectively. Table III given shows the results of false positive rate observed for three different methods, RPDEL-CRC, FRGOD and Iterative ensemble method with distance-based data filtering [2].

TABLE II. TABULATION FOR OUTLIER DETECTION ACCURACY

Samples	Outlier detection accuracy (%)		
	RPDEL-CRC	FRGOD	Iterative ensemble method with distance-based data filtering
5000	96.9	95.1	93.9
10000	95.35	94.35	91.15
15000	94.15	92.85	90.35
20000	94.05	91.55	88.15
25000	93.85	91	88
30000	93.25	90.85	87.35
35000	93	90.25	86
40000	92.55	89.85	85.25
45000	92.15	89.15	84.35
50000	92	89	83

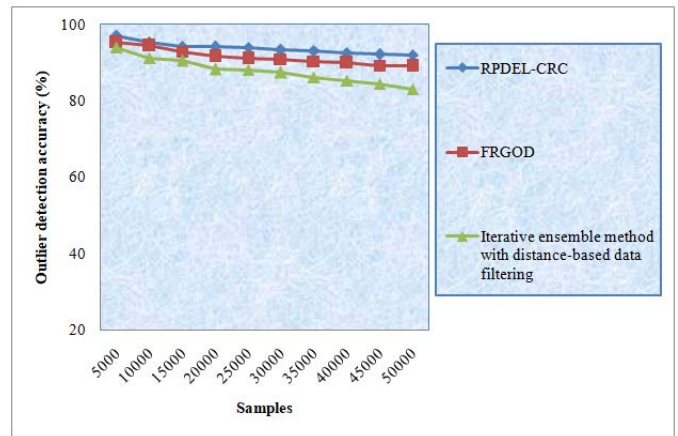


Fig. 5. Graphical Representation of Outlier Detection Accuracy.

TABLE III. TABULATION FOR FALSE POSITIVE RATE

Samples	False positive rate (%)		
	RPDEL-CRC	FRGOD	Iterative ensemble method with distance-based data filtering
5000	0.007	0.015	0.025
10000	0.015	0.018	0.026
15000	0.018	0.025	0.028
20000	0.02	0.028	0.033
25000	0.022	0.031	0.035
30000	0.025	0.032	0.036
35000	0.027	0.035	0.038
40000	0.029	0.037	0.042
45000	0.035	0.039	0.044
50000	0.038	0.042	0.048



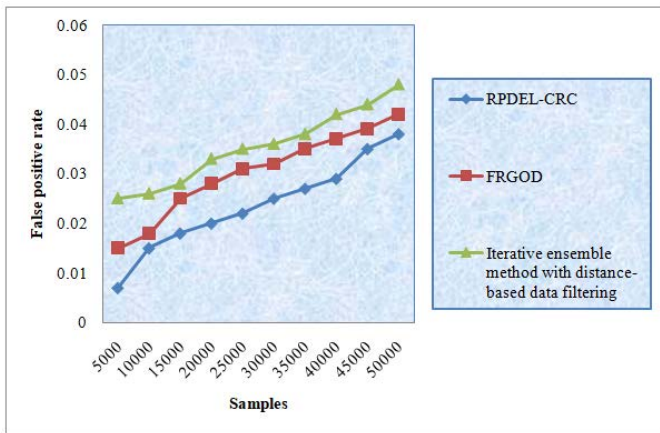


Fig. 6. Graphical Representation of False Positive Rate.

Fig. 6 given depicts false positive rate for different stock samples. From the figure, it is inferred that the false positive rate also increases with the increase in the number of stock samples involved in the simulation and hence the false positive rate is found to be directly proportional to the stock samples or samples. However, simulations conducted for 5000 samples show a false positive rate of 0.007 using RPDEL-CRC, 0.015 using FRGOD [1] and 0.025 using Iterative ensemble method with distance-based data filtering [2]. From this, it is observed that the false positive rate is comparatively lesser using RPDEL-CRC when compared to [1] and [2]. The reason behind is the application of Chebyshev Temporal and Reflective Correlation-based Outlier Detection algorithm. By applying this algorithm, dependency between two distributions of data points or data are separated. First, according to different weight of data points, i.e., based on the neighbors or data points in cluster, updated weight of a centroid data point is obtained. Next, with the identified updated weight of a centroid data point, outliers are detected based on the linear correlation between data points. Hence, by applying different updated weight of a centroid data point for each cluster, false positive rates are significantly reduced using RPDEL-CRC method by 24% compared to [1] and 36% compared to [2].

D. Case Analysis of Error Rate

Finally, the error rate involved in outlier detection is discussed in this section. The error rate is one of the significant parameters involved in the outlier detection process. This is owing to the reason that while clustering certain data points are said to be misplaced in the adjoining clusters, therefore resulting in error. This error rate is mathematically stated as given below.

$$ER = \left( \frac{V_{actual} - V_{expected}}{V_{expected}} \right) * \% \tag{18}$$

From the above (18), the error rate ‘ER’ is measured based on the actual value ‘V\_actual’ or the actual data point positioning and the expected value ‘V\_expected’ or the expected data positioning. It is measured in terms of percentage (%). Finally, Table IV lists the error rate obtained using the (18).

Finally, Fig. 7 illustrates the error rate observed during the process of outlier detection. From the figure, an increasing

trend is found to be observed using all the three methods, RPDEL-CRC, FRGOD [1] and Iterative ensemble method with distance-based data filtering [2] increasing the stock samples. This is due to the reason that with the increase in the stock samples provided as input obtained during different time instances from different companies, first, clusters are performed. While performing the clustering based on data points certain data points due to temporal instances cause a small shift in the positioning of clusters. This in turn results in the deviation and therefore error. However, simulations conducted with 5000 samples with actual data positioning observed to be 53, the expected data positioning using the three methods were observed to be 48, 45 and 43. Hence, the error rate were found to be 9.4%, 15.09% and 18.86% respectively using the three methods, therefore reducing the error with RPDEL-CRC method. The reason behind the minimization of error using RPDEL-CRC method was due to the application of Gaussian Random Projection-based Deep Extreme Learning-based Clustering algorithm. By applying this algorithm, the learning rate for solving the output weights were estimated by means of square gradient. As a result, the error rate using RPDEL-CRC was said to be reduced by 28% compared to [1] and 45% compared to [2].

TABLE IV. TABULATION FOR ERROR RATE

Samples	Error rate (%)		
	RPDEL-CRC	FRGOD	Iterative ensemble method with distance-based data filtering
5000	9.4	15.09	18.86
10000	9.75	16.15	21.32
15000	10.35	16.35	22
20000	10.85	17.25	22.85
25000	11.35	17.85	24.35
30000	13.15	18.35	24.85
35000	15.25	20	25
40000	17.35	21.35	28
45000	19.55	22.45	29.35
50000	21.25	23	30

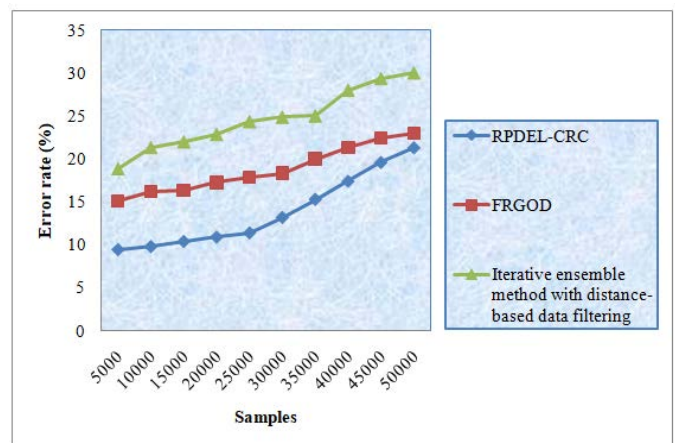


Fig. 7. Graphical Representation of Error Rate.

## VI. CONCLUSION

In spite of the fact that there has been an improvement in outlier detection, nevertheless mushrooming outliers are still found in its disastrous intents. In this day and age, it has become a big ultimatum that the behavior of outliers has to be monitored in many data mining tasks. In this paper, we proposed a new outlier detection method, called, Random Projection Deep Extreme Learning-based Chebyshev Reflective Correlation (RPDEL-CRC). The main contributions of our proposed RPDEL-CRC method is to the field of outlier detection that reduces the outlier detection time, error and false positive rate involved with maximum accuracy. The proposed method reduces the outlier detection time and error for operating the outlier detection via Gaussian Random Projection-based Deep Extreme Learning model that initially performs the clustering process by means of Gaussian Random Projection function. Next, with behavior grouping and clustering using Deep Extreme Learning, false positive rate is reduced in a timely manner. Third, the actual outlier detection process based on the clustering results is performed by means of Chebyshev Temporal and Reflective Correlation-based Outlier Detection model. Simulations were performed to evaluate the performance of RPDEL-CRC, FRGOD and Iterative ensemble method with distance-based data filtering method. Simulation results revealed that the proposed RPDEL-CRC method outperforms, FRGOD and Iterative ensemble method with distance-based data filtering method implementations, in terms of outlier detection time, accuracy, error rate and false positive rate.

### REFERENCES

- [1] Zhong Yuan, Hongmei Chen, Tianrui Li, Binbin Sang, and Shu Wang, "Outlier Detection Based on Fuzzy Rough Granules in Mixed Attribute Data," *IEEE Transactions on Cybernetics*, May 2021 [fuzzy rough granules-based outlier detection (FRGOD)].
- [2] Bodhan Chakraborty, Agneet Chatterjee, Samir Malakar, and Ram Sarkar, "An iterative approach to unsupervised outlier detection using ensemble method and distance-based data filtering," *Springer Complex & Intelligent Systems*, Feb 2022, [Iterative ensemble method with distance-based data filtering].
- [3] Asma Belhadi, Youcef Djenouri, Djamel Djenouri, Tomasz Michalak, and Jerry Chun-Wei Lin, "Deep Learning Versus Traditional Solutions for Group Trajectory Outliers," *IEEE Transactions on Cybernetics*, Dec 2020.
- [4] Saihua Cai, Ruizhi Sun, Shangbo Hao, Sicong Li, and Gang Yuan, "An Efficient Outlier Detection Approach on Weighted Data Stream Based on Minimal Rare Pattern Mi," *IEEE Xplore*, Oct 2019.
- [5] Asma Belhadi, Youcef Djenouri, Gautam Srivastava, Djamel Djenouri, Jerry Chun-Wei Lin, and Giancarlo Fortino, "Deep learning for pedestrian collective behavior analysis in smart cities: A model of group trajectory outlier detection," *Information Fusion*, Elsevier, Feb 2021.
- [6] Hongwei Du, Qiang Ye, Zhipeng Sun, Chuang Liu, and Wen Xu, "FAST-ODT: A Lightweight Outlier Detection Scheme for Categorical Data Sets," *IEEE Transactions of Network Science and Engineering*, Oct 2020].
- [7] You Lin, and Jianhui Wang, "Probabilistic Deep Autoencoder for Power System Measurement Outlier Detection and Reconstruction," *IEEE Transactions on Smart Grid*, Jul 2019.
- [8] N. Jayanthi, Burra Vijaya Babu, and N. Sambasiva Rao, "An ensemble framework based outlier detection system in high dimensional data using Tree Technique," *Materials Today: Proceedings*, Elsevier, Nov 2020.
- [9] Atiq ur Rehman, and Samir Brahim Belhaouari, "Unsupervised outlier detection in multidimensional data," *Journal of Big Data*, Springer, Jun 2021.
- [10] Mahsa Salehi, Christopher Leckie, James C. Bezdek, Tharshan Vaithianathan and Xuyun Zhang, "Fast Memory Efficient Local Outlier Detection in Data Streams," *IEEE Transactions on Knowledge and Data Engineering*, Nov 2016.
- [11] Huiping Li, BinWang, and Xin Xie, "An improved content-based outlier detection method for ICS intrusion detection," *EURASIP Journal on Wireless Communications and Networking*, Springer, Aug 2020.
- [12] Ángela Fernández, Juan Bella, and José R. Dorronsoro, "Supervised outlier detection for classification and regression," *Neurocomputing*, Springer, Feb 2022.
- [13] Zhen Cheng, En Zhu, Siqi Wang, Pei Zhang, and Wang Li, "Unsupervised Outlier Detection via Transformation Invariant Autoe," *IEEE Access*, Mar 2021.
- [14] Hongzhi Wang, Mohamed Jaward Bah, and Mohamed Hammad, "Progress in Outlier Detection Techniques: A Survey," *IEEE Access*, Aug 2019.
- [15] Xie Ling, Jia Yanlin, Xiao Jin, Gu Xin, and Huang Jing, "GMDH-Based Outlier Detection Model in Classification Problems," *Journal of Systems Science and Complexity*, Springer, Feb 2020.
- [16] Ali Degirmenci and Omer Karal, "Robust Incremental Outlier Detection Approach Based on a New Metric in Data Streams," *IEEE Access*, Nov 2021.
- [17] Fangyi Wan, Gaodeng Guo, Chunlin Zhang, Qing Goo, and Jie Liu, "Outlier Detection for Monitoring Data Using Stacked Autoencoder," *IEEE Access*, Nov 2019.
- [18] Ping Yang, Dan Wang, Zhuojun Wei, Xiaolin Du, and Tong Li, "An Outlier Detection Approach Based on Improved Self-Organizing Feature Map Clustering Algorithm," *IEEE Access*, Aug 2019.
- [19] Lu Liu, and Shang Wang, "Meta-path-based outlier detection in heterogeneous information network," *Frontiers of Computer Science*, Springer, Nov 2019.
- [20] Jintao Song, Shengfei Zhang, Fei Tong, Jie Yang, Zhiquan Zeng, and Shuai Yuan, "Outlier Detection Based on Multivariable Panel Data and K-Means Clustering for Dam Deformation Monitoring Data," *Advances in Civil Engineering*, Hindawi, Dec 2021.



# Efficient Decentralized Sharing Economy Model based on Blockchain Technology: A Case Study of Najm for Insurance Services Company

Atheer Alkhamash, Kawther Saeedi, Fatmah Baothman, Rania Anwar Aboalela, Amal Babour\*

Information Systems Department, Faculty of Computing and Information Technology  
King Abdulaziz University, Jeddah 21589, Saudi Arabia

**Abstract**—Blockchain is an emerging technology that is used to address ownership, centrality, and security issues in different fields. The blockchain technology has converted centralized applications into decentralized and distributed ones. In existing sharing economy applications, there are issues related to low efficiency and high complexity of services. However, blockchain technology can be adopted to overcome these issues by effectively opening up secure information channels of the sharing economy industry and other related parties, encouraging industry integration and improving the ability of sharing economy organizations to readily gain required information. This paper discusses blockchain technology to enhance the development of insurance services by proposing a five-layer decentralized model using Ethereum platform. The Najm for Insurance Services Company in Saudi Arabia was employed in a case study for applying the proposed model to effectively solve the issue of online underwriting, and to securely and efficiently enhance the verification and validation of transactions. The paper concludes with a review of the lessons learned and provides suggestions for blockchain application development process.

**Keywords**—Blockchain; decentralized; Ethereum; multichain; Najm; sharing economy

## I. INTRODUCTION

Blockchain is a new technology that is defined as “a distributed database, which sequentially stores a chain of data packaged into locked blocks in a safe and unchanging way” [1]. Traditionally, most of the electronic service providers are centralized entities that are required to be validated and verified, and they need to be trusted by their stakeholders. By utilizing the concept of platform cooperation with peer-to-peer (P2P) rather than centralized services, blockchain technology offers the infrastructure for decentralized security, verifiability, and trust [2]. However, several blockchain designs and implementations resulted in principally different governance structures, such as hierarchy over meritocracy, necessitating comprehensive communication among all involved stakeholders [3]. To ensure a balance in sharing economy settings, a trusted platform in centralized architectures should be replaced with blockchain technology and associated protocols to ensure trust, security, and privacy [4]. The design of such decentralized platforms based on blockchain technology and its practical implications regarding security, privacy, and trust is extremely reliant on the type of blockchain technology used [5].

Along with the advancement of blockchain technology, sharing economy is another IT-mediated development that is growing rapidly. Sharing economy can be defined as: “the sharing activity of underutilized assets with the help of IT-based technology” [6]. It is an umbrella term related to activities of sharing goods and services such as exchanging or renting them via IT, without the need to change their ownership. It enhances efficiency and effectiveness by minimizing the cost of transactions and raises the utilization and exchange of goods and services. It also enhances competition between competitors within a marketplace and minimizes the complacency of suppliers [7], [8]. Integrating blockchain into sharing economy would improve the sharing economy in terms of security and privacy, and it might increase the distribution of P2P businesses due to the provision of a high level of data integrity and nonexistence of third parties; consequently, data security would be ensured.

Therefore, in this study, a new model for integration of blockchain and sharing economy was proposed as a five-layer decentralized model and it was applied to the Najm for Insurance Services Company as a case study. The main contributions of this study are the following:

- A model for car accident report application was designed for the Najm for Insurance Services Company using a blockchain platform.
- The proposed model was developed.

The remainder of this paper is organized as follows: Section 2 discusses related works; Section 3 details existing car insurance services; Section 4 discusses the design of the proposed car insurance application based on blockchain; Section 5 covers the implementation of the proposed model; and Section 6 presents the conclusions and discusses directions for future work.

## II. RELATED WORK

### A. Blockchain Concept

Blockchain is a new technology that is operated without the need for third parties to exchange their transaction data. Blockchain can also be defined as “an appending only, ever-growing chain of blocks, which are linked sequentially using the hash pointers as a linear linked list” [9]. Specifically, as shown in Fig. 1 the block header contains a hash pointer that is linked directly to the previous block, called the parent

\*Corresponding Author.

block, and this linkage extends all the way back to the first block, called the genesis block. Further, all transactions are ordered based on Merkle trees. As a result, data on the blockchain cannot be changed unless all the subsequent blocks are altered.

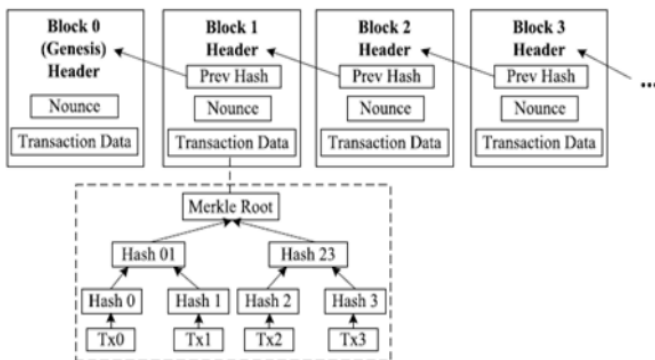


Fig. 1. Structure of Blockchain.

Blockchain can be one of three types: public blockchain, private blockchain, and consortium blockchain. The core components of the blockchain architecture are described below [10]:

- Node: the blockchain transaction that represents a user or a computer. It is the smallest construction unit of any blockchain system.
- Block: a data structure utilized to control a sequence of transactions disseminated to other network nodes.
- Chain: a series of blocks in a specific order.
- Miners: a collection of rules and agreements to conduct blockchain operations; these are unique nodes that conduct block verification processes.
- Consensus mechanism: it is called the consensus protocol and is the core of blockchain platforms. Its algorithm is used to validate the blocks, order them, and ensure agreement between nodes across the distributed P2P network. There are two main types of consensus mechanisms: proof-of-work (PoW), which is trusted and has a strong history, and proof-of-stake (PoS), which requires relatively fewer computations, less energy-consumption, and is more reliable and scalable.

The main driving features of the blockchain technology include: decentralization, transparency, security, immutability, and privacy. These are explored below [11]:

- Decentralization refers to governing the transactions to be updated from the ledger through transmission of the responsibilities to regional nodes that independently verify the transactions and then transmit them for computation with high throughput such as by employing the longest chain rule. Under decentralized consensus, no central authority or integration point is needed to receive the transactions and verify set rules. Further, no failure or trust is possible at a single point.

- Transparency refers to auditing of records by a predefined set of participants that can either be less or more open. These records are traceable and transparent, and the participants have the choice to combine their individual weighted rights.
- Security is a blockchain feature that takes the form of irreversible records stored in shared, replicated, and tamperproof ledgers. Additionally, the records cannot be forged because of the use of one-way cryptographic hash-functions. Blockchains are secure to some extent; however, security is a relative feature and having a private key allows data to be transferred throughout a blockchain.
- Immutability refers to the features of non-repudiation and irreversibility of records that control blockchain function. The data recorded in the ledger, which is tamper-resistant, cannot be secretly altered unless the alteration is done with the knowledge of the network, thereby making blockchains immutable.

Blockchain infrastructure is defined and implemented using interconnected devices in the hardware layer. The infrastructures of two major blockchain platforms are discussed below [12]:

'Ethereum' blockchain is a decentralized, open-source, generic and public blockchain network in which the transactions are validated through the PoS consensus. Its business model is used for business-to-consumer (B2C) activities. Every node has equity and can participate in creating new blocks. In contrast, the 'Hyperledger Fabric' blockchain is a decentralized, open-source, modular and private blockchain network in which the transactions are validated by different types of consensus. Its business model is used for business-to-business (B2B) activities. Further, each node in this blockchain does not have equity and thus, it cannot participate in creating new blocks.

### B. Blockchain-related Application

Blockchains may be used in different fields. The following are some examples where blockchain is used to enhance different functionalities [13]:

Gem is a Central Disease Control system that uses blockchain to provide effective disaster relief by inputting outbreak information onto a blockchain having an immutable ledger to help researchers collect and analyze data, and make the circumstances of the disease spread evident. The Ethereum platform has been used in Gem [14].

Abra is a cryptocurrency wallet that employs the Bitcoin blockchain network to control balances saved in different regions. It operates an interest-earning service for stable coins and cryptocurrencies, a trading service for buying and selling cryptocurrencies, and lending services. The Ethereum platform has been used in Abra [15].

Augur is a market prediction system that uses blockchain technology for financial services and it is based on a decentralized ecosystem. The main purpose of creating the Augur platform was to serve as a warning system in all matters. It also allows people to invest and profit through their

expectations and opinions, which enables Augur to provide more accurate forecasting of upcoming events. The Ethereum platform has been used in Augur [16].

Skuchain is a control goods and services system based on blockchain that is used to track the services provided through a supply chain. It assists businesses in overcoming the difficulties and expenses associated with their inventory and help to advance the concept of "collaborative commerce." The Hyperledger platform has been used in Skuchain [17].

OpenBazaar was a blockchain-based decentralized trading market where no middlemen were needed to trade goods and services. It also utilized Bitcoin, which is a censorship-resistant, decentralized, and inexpensive digital currency. The Ethereum platform has been used in OpenBazaar [18].

Blockchain is used to enhance the functionality of car reporting systems in different ways. For instance, the blockchain-based reporting system proposed in [19] provides a new mechanism to provide indisputable accident forensics by guaranteeing verifiability and trustworthiness of information. The proposed mechanism is used to verify and validate a new block of event data in a trusted manner without any central ownership. In addition, the model proposed in [20], called (IV-TP), is used to communicate between intelligent vehicles using blockchain technology. The data shared between the intelligent vehicles are secure and reliable. The (IV-TP) model provides trustworthiness for vehicles' histories (i.e., behavior) and their actions (legal or illegal). Blockchain is used to store all the details of each vehicle.

The present paper proposes an architecture of a car accident reporting application for the Najm for Insurance Services Company using the blockchain platform. Najm is a Saudi company started in 2007 to implement an effective platform for facilitating, overcoming, and resolving accident-related transactions and formalities [21]. This organization has obliged itself to offer hassle-free and smooth processing among insurance companies. Najm initially had only 13 insurance agencies as its core shareholders but now, it works with more than 26 Saudi insurance companies.

### III. EXISTING CAR INSURANCE SERVICES

This section demonstrates the existing accident reporting systems for the Najm for Insurance Services Company [21].

As shown in Fig. 2, the process begins when the insured party first records the accident on an application or contacts the car accident reporting company. Second, the company appoints a surveyor located close to the accident spot. Third, the surveyor arrives at the accident spot. Fourth, the surveyor examines and evaluates the accident. Fifth, the insured party receives a report from the surveyor. Sixth, the insured party goes to the traffic police to get a repair paper in the report stamped. Seventh, the insured party gets the estimated damage cost from the authorized assessment Center. Eighth, the insured party goes to the insurance company. Finally, the insured party obtains a final clearance or written approval from the insurance company to get the vehicle repaired.

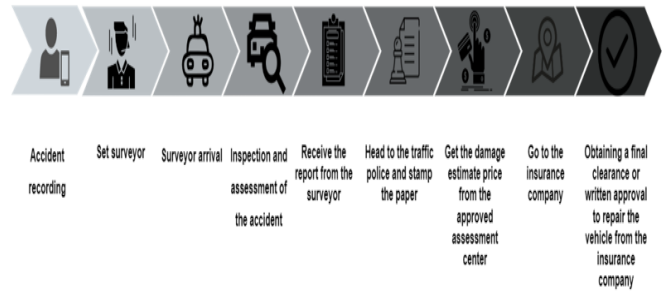


Fig. 2. Existing Accident Procedure in the Najm for Insurance Services Company.

Therefore, the existing accident reporting systems such as that of the Najm for Insurance Services Company, are suffering from various challenges, including the need to enhance user security, increase service availability, and provide service flexibility at the lowest cost. Moreover, the accident report data must not be changed after they are inputted into the database. Further, accident reports should be shared and distributed among different stakeholders such as the insured party, reporting company, surveyor, traffic police, assessment centers, and insurance company in a consistent and secure manner. These issues should be addressed to provide a trusted accident reporting system.

### IV. DESIGN OF THE PROPOSED CAR INSURANCE APPLICATION BASED ON BLOCKCHAIN

After conducting a thorough synthetic analysis about the use of blockchain technology in sharing economy applications, this study combined the blockchain technology with sharing economy applications to create a decentralized P2P sharing economy model for a car accident reporting application. In general, the sharing economy for businesses connects service providers and consumers, and enables people to offer services through online platforms in a private and secure manner. As mentioned hereinabove, the blockchain technology facilitates decentralization of data storage and communication. The decentralized processes can be automated when deploying 'smart contracts' that are a part of the blockchain technology.

To build a decentralized sharing economy using the blockchain technology, several design elements should be considered, including the features of blockchain and its main components, as well as decentralized infrastructures and platforms. As previously stated, the blockchain architecture presents benefits for businesses in gaining several features such as improved security and privacy.

The proposed model is adapted from the 'Blockchain Market Engineering Framework' [22]. The Ethereum platform was selected for implementation in the proposed model because it is an open-source, public, and generic blockchain architecture in which the transactions are validated by the PoS consensus. The proposed blockchain architecture is independent of any specific platforms and it consists of five main layers, as shown in Fig. 3.

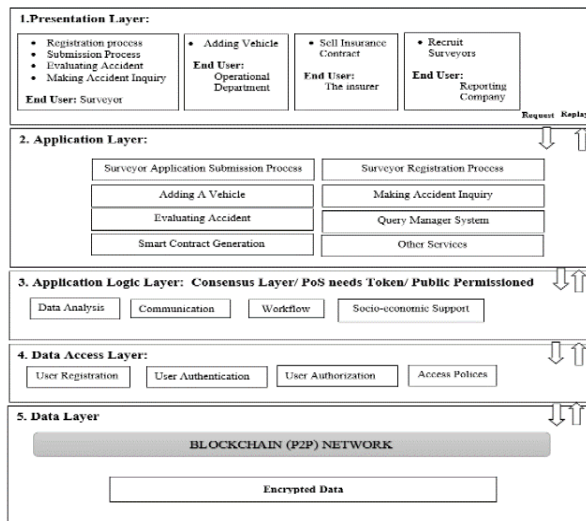


Fig. 3. Five-layer Architecture of the Proposed Model.

The blockchain platform (BaaS) provider manages the setup, maintenance, and support of the blockchain infrastructure. A node is any electronic equipment that is linked to the internet and has an IP address, such as a computer, mobile phone, or printer. The node can copy the blockchain and process transactions.

The data layer manages encrypted data that are maintained in a standard form to ensure compatibility and inter-sectoral communication requirements.

The data access layer integrates the data generated from various sources to enable user registration, user authentication, user authorization and accessing of policies. Effective access control is required to ensure that only the involved actors (i.e., legal users) have access to the relevant transaction data. The implementation of a role concept is critical in this context. Furthermore, a precisely designed role concept combined with effective access control makes illegal data alteration more difficult. Furthermore, the integration of related organizational parties and relevant metadata is supported by a metadata repository.

The application logic layer manages data analysis, communication, socioeconomic support, and workflow of the integrated applications and systems. Therefore, the processed and retained data are available for all purposes. This predominantly concerns research trends and allows to conduct statistical assessments.

The application layer represents the platform features, business services and user-controlled services that are given to the users of sharing economy organisations, such as the surveyor application submission process, surveyor registration process, addition of a vehicle, making an accident inquiry, evaluation of an accident, querying the manager system, smart contract generation, and other services.

Lastly, the presentation layer enables the services provided by the system to be displayed to stakeholders such as the surveyor, insurer, reporting company and operational department. The functions and features provided to each user differ depending on their role and permission. For instance,

the registration process, submission process, accident evaluation, and making an accident inquiry are available for the surveyor; addition of vehicles is available to the operational department; selling of insurance contracts is available to the insurers; and recruitment of surveyors is available to the reporting company. Note that it is vital to ensure responsiveness in the proposed model by allowing access through web or desktop applications supported by various operating systems. Also, the model can be further extended by applying security mechanisms in smartphones, such as fingerprint and iris scanning.

To demonstrate the effectiveness of the proposed model, it was adapted in the Najm for Insurance Services Company [21].

## V. IMPLEMENTATION OF THE PROPOSED APPLICATION

To implement the prototype, the following steps were undertaken:

- Creating a blockchain for data storing: A multichain platform was used with the Windows 10 OS; the chain consisted of only one computer and all data about stakeholders were saved in the chain as bytecode that was then converted to the JSON format to be read. Further, any stakeholder computer could be added, and the command prompt (CMD) scripts were used to create the chain.
- Connecting with the chain: To connect with the chain and cover its complexity, an API was created using the Python programming language. The created API offers data storage and retrieval from the chain. Further, the Flask web server was used to allow a client to connect to the chain.
- The application: To create a GUI app for users, an android app was used to connect with the API.

### A. Code and Functions for Computer Program

The system consists of a mobile application and a computer program. The computer program consists of two elements: one is responsible for the blockchain and the other is responsible for the web server. The blockchain and web server functions are listed in Tables I and II, respectively.

### B. Proposed Mobile Application

The proposed application contains different screens for executing different functionalities. Fig. 4 represents the pre-step of using the system. The user needs to enter their IP address to link the application to the blockchain and then, select one of three options: 'Create New User', 'Login', or 'Report an Accident'.

The surveyor is provided two options: 'Create New User' and 'Login'.

Create New User: When the surveyor chooses to create an account, the screen contains fields to receive information about the user. The surveyor should enter their name, email, password, form number, car plate, identification number and mobile number and then, click on the 'Create User' icon, as shown in Fig. 5.

TABLE I. BLOCKCHAIN FUNCTIONS

Functions	Description
Create	It is used to create a new blockchain.
Start	It is used to start a blockchain activity.
Storage	It is used to create a stream on blockchain to start writing on it and give permission for commuters who are connected to the blockchain to write on it.
add_data	It is used to take (chain, stream, path) of blockchain to write on it and publish.
converter	It is used to take the existing data and convert it to hexadecimal.
find_data	It is used to find data inside a blockchain.
distance	It is used to calculate the distance between the surveyor and the insured party.
find_ac	It is used to store all the data in the JSON file and search for the existing data.

TABLE II. WEB SERVER FUNCTIONS

Functions	Description
app.route(users)	It is used to add a new user to the company.
app.route(acid)	It is used to report an accident.
app.route(Log)	It is used to log in to the user account and locate the nearest accident.
app.route(Select)	It is used to the nearest accident for serving.



Fig. 4. Initial Registration.



Fig. 5. Create an Account.

After the surveyor account creation process is completed, the screen shown in Fig. 6 is presented to the surveyor, which contains the message “Your Req Has been sent, wait for a phone call”.

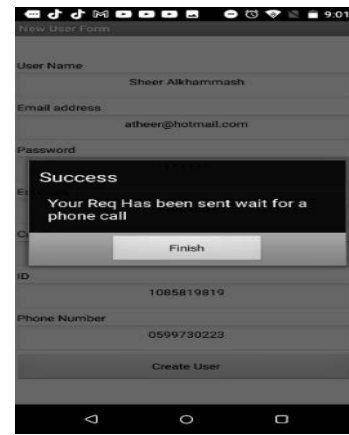


Fig. 6. Request Message.

Login: When the surveyor selects the “Login” option, they will be allowed to login. This screen contains two fields: email and password. After providing this information, the surveyor is logged into the application, as shown in Fig. 7.



Fig. 7. Login.

After the surveyor login process is successfully executed, the screen shown in Fig. 8 appears which contains longitude and latitude coordinates of the nearest accidents.



Fig. 8. Nearest Accidents.

On the other hand, the insurer has one option:

Report an Accident: When the insurer chooses the third option, a screen appears containing information about the



accident, mobile number of the insured party, car plate number, username, and accident degree. There are three different accident degrees: "small no injury", "medium no injury", and "With injury", as shown in Fig. 9 and 10.



Fig. 9. Report an Accident.

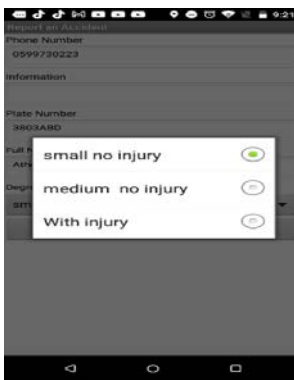


Fig. 10. Accident Degree.

After determining the accident degree, the insurer presses submit and the screen shown in Fig. 11 appears for the insurer, which contains the message "Accident Data Has been Received".

CMD scripts are used for interacting with blockchain. Fig. 12 shows how the user information is saved in blockchain. Fig. 13 shows how the accident report is stored in blockchain.

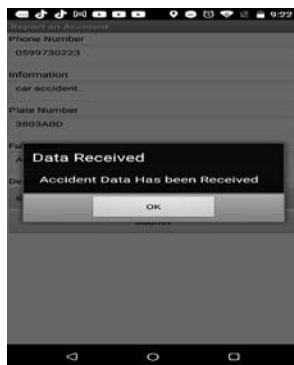


Fig. 11. Submit Accident Report Form.

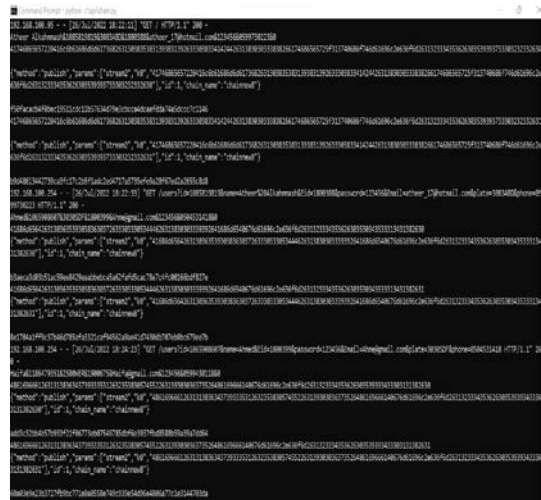


Fig. 12. User Information.

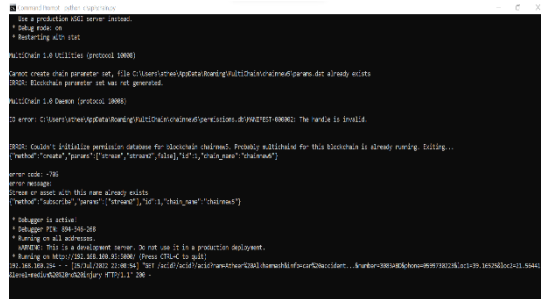


Fig. 13. Accident Reporting Information.

## VI. CONCLUSION AND FUTURE WORK

Blockchain plays an important role in increasing data security and integrity in sharing economy applications. Therefore, proposing a blockchain-based framework to enhance sharing economy application is important to solve the issues related to existing accident reporting systems, including the need for significant data harvesting, providing secure storage, and enabling frequently updated transactions. Moreover, the old reporting data must not be changed after inputting it into the database. Also, accident reports should be shared and distributed among different stakeholders in a consistent and secure manner. These issues should be addressed to provide a trusted accident reporting system. In this paper, a new business method is proposed for sharing economy using the blockchain architecture. The proposed model features a software architecture that was used to develop the sharing economy application based on blockchain and the identified business pattern and architectural model of a car accident reporting application was adopted as a prototype of the proposed blockchain platform and subsequently implemented for the Najm for Insurance Services Company in Saudi. With regards to future work, we need to add new functions and evaluate the adaptation of the architecture and business pattern identified through the case study.

### REFERENCES

[1] M. Turkanović, M. Hölbl, K. Košič, M. Heričko, and A. Kamišalić, "EduCTX: A blockchain-based higher education credit platform," IEEE access, 2018, vol. 6, pp. 5112–5127, doi: 10.1109/ACCESS.2018.2789929.



- [2] S. B. Taieb and R. J. Hyndman, "A gradient boosting approach to the Kaggle load forecasting competition," *International journal of forecasting*, 2014, vol. 30, no. 2, pp. 382–394, doi:10.1016/j.ijforecast.2013.07.005.
- [3] L. Hetmank, "Components and functions of crowdsourcing systems—a systematic literature review," 2013.
- [4] J. Klinger and M. Lease, "Enabling trust in crowd labor relations through identity sharing," *Proceedings of the American Society for Information Science and Technology*, 2011, vol. 48, no. 1, pp. 1–4, doi: 10.1002/meet.2011.14504801257.
- [5] H. R. Andrian and N. B. Kurniawan, "Blockchain Technology and Implementation: A Systematic Literature Review," in *2018 International Conference on Information Technology Systems and Innovation (ICITSI)*, 2018 IEEE, pp. 370–374, doi: 10.1109/ICITSI.2018.8695939.
- [6] G. Görög, "The Definitions of Sharing Economy: A Systematic Literature Review," *Management (18544223)*, 2018, vol. 13, no. 2.
- [7] P. Goudin, "The cost of non-Europe in the sharing economy: Economic, social and legal challenges and opportunities," 2016.
- [8] A. Hira and K. Reilly, "The emergence of the sharing economy: Implications for development," *Journal of Developing Societies*, 2017, vol. 33, no. 2, pp. 175–190, doi: 10.1177/0169796X17710071.
- [9] Z. Liu et al., "A survey on blockchain: A game theoretical perspective," *IEEE Access*, 2019, vol. 7, pp. 47615–47643, doi: 10.1109/ACCESS.2019.2909924.
- [10] P. Fraga-Lamas and T. M. Fernández-Caramés, "A review on blockchain technologies for an advanced and cyber-resilient automotive industry," *IEEE access*, 2019, vol. 7, pp. 17578–17598, doi: 10.1109/ACCESS.2019.2895302.
- [11] P. Tasca and C. J. Tessone, "Taxonomy of blockchain technologies. Principles of identification and classification," 2017, arXiv preprint arXiv:1708.04872, doi:10.48550/arXiv.1708.04872.
- [12] M. Valenta and P. Sandner, "Comparison of ethereum, hyperledger fabric and corda," *Frankfurt School Blockchain Center*, 2017, vol. 8, pp. 1–8.
- [13] J. J. Sikorski, J. Haughton, and M. Kraft, "Blockchain technology in the chemical industry: Machine-to-machine electricity market," *Applied energy*, 2017, vol. 195, pp. 234–246, doi: 10.1016/j.apenergy.2017.03.039.
- [14] J. Gong and L. Zhao, "Blockchain application in healthcare service mode based on Health Data Bank," *Frontiers of engineering management*, 2020, vol. 7, no. 4, pp. 605–614, doi: 10.1007/s42524-020-0138-9.
- [15] H. Natarajan, S. Krause, and H. Gradstein, "Distributed ledger technology and blockchain," 2017.
- [16] R. Hobeck, C. Klinkmüller, H. Bandara, I. Weber, and W. M. van der Aalst, "Process mining on blockchain data: a case study of Augur," in *International conference on business process management*, 2021 Springer, pp. 306–323, doi: 10.1007/978-3-030-85469-0\_20.
- [17] A. Rijanto, "Blockchain technology adoption in supply chain finance," *Journal of Theoretical and Applied Electronic Commerce Research*, 2021, vol. 16, no. 7, pp. 3078–3098, doi: 10.3390/jtaer16070168.
- [18] J. Mattila, "The blockchain phenomenon," *Berkeley Roundtable of the International Economy*, 2016, vol. 16.
- [19] H. Guo, E. Meamari, and C.-C. Shen, "Blockchain-inspired event recording system for autonomous vehicles," in *2018 1st IEEE international conference on hot information-centric networking (HotICN)*, 2018 IEEE, pp. 218–222, doi:10.1109/HOTICN.2018.8606016.
- [20] Q. Ren, K. L. Man, M. Li, B. Gao, and J. Ma, "Intelligent design and implementation of blockchain and Internet of things-based traffic system," *International Journal of Distributed Sensor Networks*, 2019, vol. 15, no. 8, doi: 10.1177/1550147719870653.
- [21] "Najm for Insurance Services Company." Available at: <https://www.najm.sa/en>.
- [22] B. Notheisen, F. Hawlitschek, and C. Weinhardt, "Breaking down the blockchain hype—towards a blockchain market engineering approach," 2017.

# Virtual Communities of Practice to Promote Digital Agriculturists' Learning Competencies and Learning Engagement: Conceptual Framework

Maneerat Manyuen<sup>1</sup>, Surapon Boonlue<sup>2\*</sup>, Jariya Nanchaleay<sup>3</sup>, Vitsanu Nittayathamkul<sup>4</sup>

Division of Learning Innovation and Technology, Faculty of Industrial Education and Technology  
King Mongkut's University of Technology Thonburi, Bangkok, Thailand<sup>1, 3</sup>

Department of Educational Communications and Technology, Faculty of Industrial Education and Technology  
King Mongkut's University of Technology Thonburi, Bangkok, Thailand<sup>2</sup>

Program in Learning Innovation, Faculty of Industrial Education  
Rajamangala University of Technology Suvarnabhumi, Suphanburi, Thailand<sup>4</sup>

**Abstract**—Virtual Communities of Practice (VCoPs) are networks of people who share a common interest and a desire to learn together in the same domain via ICT. The limitations of the existing concepts for developing VCoPs in general contexts are not explained in terms of the integration between virtual learning technologies and digital learning strategies used to promote expected learning outcomes in the agricultural sector. This research aims to propose the conceptual framework for developing the Virtual Communities of Practice through Digital Inquiry (VCoPs-DI Model) to promote digital agriculturists' learning competencies and engagement. The research methodology was divided into three stages: the first stage involves a literature review for document analysis and synthesis, the second stage involves constructing the conceptual framework, and the third stage involves evaluating the content validity index. The key results showed that the developed conceptual framework has three parts: (1) The fundamentals of concept formation were divided into four concept bases: (1.1) Communities of Practice (CoPs), (1.2) Virtual Learning Environments (VLEs), (1.3) Digital Learning Resources (DLRs), and (1.4) Critical Inquiry Method; (2) The identification of the manipulated variable was divided into two compositions: (2.1) VCoPs and (2.2) Digital Inquiry (DI); (3) The identification of the dependent variable was divided into two compositions: (3.1) Digital agriculturists' learning competencies, and (3.2) learning engagement. Findings from an expert's review show that the scale levels of the content validity index (SCVI) were 0.958. We anticipate that our conceptual framework could be used for reference as part of the design and development of the VCoPs model to promote learning in the agricultural sector.

**Keywords**—VCoPs; digital inquiry; digital agriculture; learning competencies; learning engagement

## I. INTRODUCTION

Many global issues, such as climate change, lack of natural resources, demographics, and food waste, are putting pressure on the overall sustainability of agricultural systems. In this digital age, traditional agricultural management approaches need to be radically transformed so that smart technologies can contribute to innovation and redesign the entire value chain to maintain the sustainability of the agricultural sector. Current advances in advanced digital technology tend to lead to the

fourth phase of the revolution in the agricultural sector, known as "Agriculture 4.0" [1].

Agriculture 4.0 in developing countries is characterized by low technological levels of technology. The main reasons for this are the high cost of advanced digital technology (e.g., 5G, cloud computing, Internet of Things, blockchain, data mining, artificial intelligence, augmented reality, virtual reality, etc.) and the dynamics of today's business environment. However, these types of systems are becoming increasingly important, especially for achieving the Sustainable Development Goals (SDGs), which are related to three specific goals: Zero Hunger (Goal 2), Clean Water and Sanitation (Goal 6), and Life on Land (Goal 15). One alternative to achieving this goal is to adopt Logistics 4.0 technologies and educational technology. This is because these technologies can address issues such as nutrition, food safety, and soil and water conservation [2].

Information and communication technologies (ICT), such as the Internet, e-mail, and videoconferencing, have made the human learning process more efficient and productive in terms of daily operations. However, ICT not only improves people's daily productivity but also supports the ability to share that information and tacit knowledge with both internal and external organizations as well as social networks. One of the most popular ways to share large amounts of information and tacit knowledge is through an informal learning environment such as a Community of Practices (CoPs) [3] or Virtual Community of Practices (VCoPs) [4].

Digital learning resources are characterized by being virtual spaces that learners can access through information and communication technology. Managing learning resources in a virtual space requires consideration of the media exposure and lifestyle of digital learners. Digital learning resources can be classified into five categories [5–7]: 1) search engines and translation tools; 2) data and storage management tools; 3) content creation, presentation, and publishing tools; 4) distance learning tools; 5) social networking tools.

The limitations of existing concepts for the development of VCoPs in general contexts are not explained concerning the integration between virtual learning technologies and digital

\*Corresponding Author.

learning strategies used to promote expected learning outcomes in the agricultural sector for digital learners of all ages, regardless of their background and geographical location [3–4]. Thus, this research paper focused on proposing the conceptual framework for the development of the Virtual Communities of Practice through Digital Inquiry (VCoPs-DI Model) to promote digital agriculturists’ learning competencies and learning engagement that can be used for reference in the design and development of the VCoPs model by integration between virtual learning technologies and digital learning strategies used to promote learning in the agricultural sector in the future.

## II. RESEARCH OBJECTIVES

1. To synthesize the conceptual framework for the development of the Virtual Communities of Practice through Digital Inquiry (VCoPs-DI Model) to promote digital agriculturists’ learning competencies and learning engagement.
2. To construct the conceptual framework for the development of the VCoPs-DI Model to promote digital agriculturists’ learning competencies and learning engagement.
3. To validate the conceptual framework for the development of the VCoPs-DI Model to promote digital agriculturists’ learning competencies and learning engagement.

## III. RESEARCH METHODOLOGY

Stage I: Synthesis of the conceptual framework for the development of the VCoPs-DI Model to promote digital agriculturists’ learning competencies and learning engagement. This stage was a qualitative research method based on a literature review. The researchers conducted studies, analyzed, and synthesized research papers that included Communities of Practice (CoPs), Virtual Learning Environments (VLEs), Digital Learning Resources (DLRs), the Critical Inquiry Method, Digital Agriculturists’ learning competencies, and learning engagement from ERIC, Scopus, and Web of Science online databases published during 2015–2021 using content analysis [8] of text data in research papers.

Stage II: Construct the conceptual framework [9] for the development of the VCoPs-DI Model to promote digital agriculturists’ learning competencies and learning engagement was developed by conducting studies, analysis, and synthesis through content analysis in stage I.

Stage III: Validation of the conceptual framework for the development of the VCoPs-DI Model to promote digital farmers’ learning competencies and learning engagement was an expert judgement by five educational technology experts on a four-point Likert scale for the Content Validity Index (CVI) [10] The scale used in this study was used and responses included 1 = Not Relevant, 2 = Somewhat Relevant, 3 = Relevant, and 4 = Highly Relevant. Researchers recommend that a scale with excellent content validity should be composed of I-CVIs of 0.78 or higher and S-CVI/UA and S-CVI/Ave of 0.8 and 0.9 or higher, respectively.

## IV. RESEARCH FINDINGS

### A. The First Stage

- The results of the synthesis of components of Communities of Practice (CoPs) as shown in Table I [11-18]:

TABLE I. RESULTS OF THE SYNTHESIS OF COMPONENTS OF CoPs

Components of CoPs	Reference							
	[11]	[12]	[13]	[14]	[15]	[16]	[17]	[18]
1. Purpose/ Vision/ Mission	✓	✓	✓	✓	✓	✓	✓	✓
2. Participant/ Members/ People	✓	✓	✓	✓	✓	✓	✓	✓
3. Perspective/ Paradigm/ Shared values/ Negotiation	✓	✓	✓	✓	✓	✓	✓	✓
4. Processes/ Methods/ Activity	✓	✓	✓	✓	✓	✓	✓	✓
5. Platforms/ Technology	✓	✓	✓	✓	✓	✓	✓	✓
6. Products/ Productivity	✓	✓	✓	✓	✓	✓	✓	✓

According to the result of the synthesis indicated in Table I, it can be summarized that the CoPs comprise six components: 1) purpose, 2) participants, 3) perspective, 4) processes, 5) platforms, and 6) products.

- The results of the synthesis of components of Virtual Learning Environments (VLEs) as shown in Table II [19-26]:

TABLE II. RESULTS OF THE SYNTHESIS OF COMPONENTS OF VLEs

Components of VLEs	Reference							
	[19]	[20]	[21]	[22]	[23]	[24]	[25]	[26]
1. Learning aims and objectives	✓	✓	✓	✓	✓	✓	✓	✓
2. Learning contracts			✓	✓	✓	✓		
3. Learning issues/ Topic / Agenda	✓	✓	✓	✓	✓	✓	✓	✓
4. Learning activities/ Task	✓	✓	✓	✓	✓	✓	✓	✓
5. Learning technologies	✓	✓	✓	✓	✓	✓	✓	✓
6. Learning assessments	✓	✓	✓	✓	✓	✓	✓	✓

According to the result of the synthesis indicated in Table II, it can be summarized that the VLEs comprises six components: 1) learning aims, 2) learning contracts, 3) learning issues, 4) learning activities, 5) learning technology and 6) learning assessments.

- The results of the synthesis of components of Digital Learning Resources (DLRs) as shown in Table III [5, 7, 27-29]:

TABLE III. RESULTS OF THE SYNTHESIS OF COMPONENTS OF DLRs

Components of DLRs	Reference				
	[5]	[7]	[27]	[28]	[29]
1. Search engine and translation tools	✓	✓	✓	✓	✓
2. Data and storage management tools	✓	✓	✓	✓	✓
3. Content creation, presentation, and publishing tools	✓	✓	✓	✓	✓
4. Distance learning tools	✓		✓	✓	✓
5. Social networking tools	✓	✓	✓	✓	✓

According to the result of the synthesis indicated in Table III, it can be summarized that the DLRs, comprise five components: 1) search engine and translation tools, 2) data and storage management tools, 3) content creation, presentation, and publishing tools 4) distance learning tools, and 5) social networking tools.

- The results of the synthesis of components of the critical inquiry method as shown in Table IV [30-32]:

TABLE IV. RESULTS OF THE SYNTHESIS OF COMPONENTS OF CRITICAL INQUIRY METHOD

Components of critical inquiry method	Reference		
	[30]	[31]	[32]
1. Observing/ Exploring and questioning	✓	✓	✓
2. Information seeking/ Information searching/ Problem-posing	✓		✓
3. Knowledge building/ Knowledge gathering/ Knowledge construction/Taking action	✓	✓	✓
4. Creative communicating	✓	✓	✓
5. Knowledge sharing/ Knowledge exchange	✓	✓	✓

According to the result of the synthesis indicated in Table IV, it can be summarized that the critical inquiry method comprises five components: 1) exploring and questioning, 2) information searching, 3) knowledge building, 4) creative communicating and 5) knowledge sharing.

- The results of the synthesis of components of the digital agriculturists' learning competencies as shown in Table V [33-38].

According to the result of the synthesis indicated in Table V, it can be summarized that the digital agriculturists' learning competencies comprise six components: 1) seeking lifelong learning opportunities, 2) self-concept of being an effective digital learner, 3) initiative, creativity, and independent learning concerning digital agriculture learning issues, 4) self-responsibility in digital agriculture occupations, 5) optimistic about agriculture's evolution in the digital era, 6) problem-solving and decision-making concerning agriculture practices in the digital era.

TABLE V. RESULTS OF THE SYNTHESIS OF COMPONENTS OF DIGITAL AGRICULTURISTS' LEARNING COMPETENCIES

Components of Digital agriculturists' learning competencies	Reference
1. Seeking lifelong learning opportunities	[33], [34], [35], [36], [37]
2. Self-concept of being an effective digital learner	[33], [34], [36]
3. Initiative, creativity and independent learning concerning digital agriculture learning issues	[33], [34], [35], [38]
4. Self-responsibility in digital agriculture occupations	[33], [34], [35], [36], [37], [38]
5. Optimistic about agriculture evolution in the digital era	[33], [34], [35], [36], [37], [38]
6. Problem solving and decision-making concerning agriculture practices in the digital era.	[33], [34], [35], [36], [37]

- The results of the synthesis of components of the learning engagement as shown in Table VI [39-41]:

TABLE VI. RESULTS OF THE SYNTHESIS OF COMPONENTS OF LEARNING ENGAGEMENT

Components of learning engagement	Reference		
	[39]	[40]	[41]
1. Behavioral engagement	✓	✓	✓
2. Emotional engagement	✓	✓	✓
3. Cognitive engagement	✓	✓	✓

According to the result of the synthesis indicated in Table VI, it can be summarized that learning engagement comprises three components: 1) behavioral engagement, 2) emotional engagement, 3) cognitive engagement.

*B. The Second Stage*

- The results of the construction of the conceptual framework for the development of the Virtual Communities of Practice through Digital Inquiry (VCoPs-DI Model) to promote digital agriculturists' learning competencies and learning engagement are in Fig. 1.

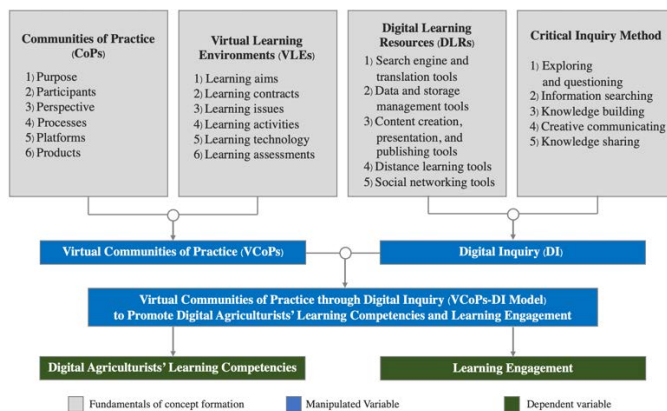


Fig. 1. Conceptual Framework.

From Fig. 1, this conceptual framework consists of the following three parts: (1) The fundamentals of concept formation were divided into four concept bases: a) Communities of Practice (CoPs), b) Virtual Learning Environments (VLEs), c) Digital Learning Resources (DLRs), and d) Critical Inquiry Method. (2) The identification of the manipulated variable was divided into two compositions: a) Virtual Communities of Practice (VCoPs), which has six elements: i) desire and passion for learning, ii) digital learning platforms, iii) decentralized learning contracts, iv) deep learning issues of practice, v) dynamic learning processes, vi) developmental learning assessments and b) Digital Inquiry (DI), which has five elements: i) exploring and questioning, ii) information searching, iii) knowledge building, iv) creative communicating, v) knowledge sharing. (3) The identification of the dependent variable was divided into two compositions: a) Digital agriculturists' learning competencies, and b) Learning engagement.

C. The Third Stage

The results of the validation of the conceptual framework for the development of the VCoPs-DI Model to promote digital agriculturists' learning competencies.

TABLE VII. RESULTS OF THE VALIDATION OF THE CONCEPTUAL FRAMEWORK FOR THE DEVELOPMENT OF THE VCoPs-DI MODEL TO PROMOTE DIGITAL AGRICULTURISTS' LEARNING COMPETENCIES

Evaluation Items	Experts					Number in Agreement	I-CVI	Meaning
	1	2	3	4	5			
<b>Part I: Fundamentals of concept formation</b>								
1. Communities of Practice (CoPs)	4	4	4	4	4	5	1.00	Valid
2. Virtual Learning Environments (VLEs)	4	4	4	4	4	5	1.00	Valid
3. Digital Learning Resources (DLRs)	4	4	4	4	4	5	1.00	Valid
4. Critical Inquiry Method	2	4	4	4	4	4	0.80	Valid
<b>Part II: Identification of manipulated variable</b>								
<b>1. Virtual Communities of Practice (VCoPs)</b>								
1) Desire and passion for learning	4	4	4	4	2	4	0.80	Valid
2) Digital learning platforms	3	4	4	4	4	5	1.00	Valid
3) Decentralized learning contracts	3	3	4	3	4	5	1.00	Valid
4) Deep learning issues of practice	3	3	4	3	4	5	1.00	Valid
5) Dynamic learning processes	3	3	3	3	4	5	1.00	Valid
6) Developmental learning assessments	2	3	4	3	4	4	0.80	Valid
<b>2. Digital Inquiry (DI)</b>								
1) Exploring and	4	4	4	4	4	5	1.00	Valid

questioning									
2) Information searching	4	4	4	4	4	5	1.00	Valid	
3) Knowledge building	3	4	4	4	4	5	1.00	Valid	
4) Creative communicating	3	4	4	4	4	5	1.00	Valid	
5) Knowledge sharing	4	4	4	4	4	5	1.00	Valid	
<b>Part III: Identification of dependent variable</b>									
<b>1. Digital agriculturists' learning competencies</b>									
1) Seeking lifelong learning opportunities	4	4	4	4	4	5	1.00	Valid	
2) Self-concept of being an effective digital learner	4	4	4	4	4	5	1.00	Valid	
3) Initiative, creativity and independent learning concerning digital agriculture learning issue	3	2	4	4	4	4	0.80	Valid	
4) Self-responsibility in digital agriculture occupations	4	4	4	4	4	5	1.00	Valid	
5) Optimistic about agriculture evolution in the digital era	4	4	4	4	4	5	1.00	Valid	
6) Problem-solving and decision-making concerning agriculture practices in the digital era	3	4	3	4	3	5	1.00	Valid	
<b>2. Learning engagement</b>									
1) Behavioral engagement	4	4	4	4	4	5	1.00	Valid	
2) Emotional engagement	4	4	4	4	4	5	1.00	Valid	
3) Cognitive engagement	4	4	4	4	4	5	1.00	Valid	
<b>S-CVI/Ave</b>							0.958	Excellent content validity	

From Table VII, based on the experts' review, the results of the validation of the conceptual framework for the development of the VCoPs-DI Model to promote digital agriculturists' learning competencies found that the item levels of the content validity index (I-CVI) were in the range of 0.80–1.00.

In addition, the scale levels of the content validity index (SCVI) also showed excellent content validity (S-CVI/Ave ≥ 0.90).

## V. CONCLUSION AND DISCUSSION

The proposed conceptual framework for developing the Virtual Communities of Practice through Digital Inquiry (VCoPs-DI Model) to promote digital agriculturists' learning competencies and learning engagement has three parts as follows:

1. The fundamentals of concept formation are divided into four concept bases: *a*) Communities of Practice (CoPs), *b*) Virtual Learning Environments (VLEs), *c*) Digital Learning Resources (DLRs), and *d*) Critical Inquiry Method.

2. The identification of the manipulated variable was divided into two compositions: *a*) Virtual Communities of Practice (VCoPs), which has six elements: *i*) desire and passion for learning, *ii*) digital learning platforms, *iii*) decentralized learning contracts, *iv*) deep learning issues of practice, *v*) dynamic learning processes, *vi*) developmental learning assessments and *b*) Digital Inquiry (DI), which has five elements: *i*) exploring and questioning, *ii*) information searching, *iii*) knowledge building, *iv*) creative communicating, *v*) knowledge sharing.

3. The identification of the dependent variable was divided into two compositions: *a*) digital agriculturists' learning competencies and *b*) learning engagement. The five educational technology experts evaluated the validity of the conceptual framework for developing the VCoPs-DI Model to promote digital agriculturists' learning competencies and found that the validity of the proposed conceptual framework has excellent content validity.

From the findings, we anticipate that our conceptual framework could be used for reference in the design and development of the VCoPs model to promote learning in the agricultural sector. We expect that our framework will lead to practical ways to incorporate agricultural communication to increase participation in VCoPs among the agricultural workforce, including for all ages in the digital transmedia era, towards the Digital Agriculturists' learning competencies, which include six expected learning outcomes: 1) Seeking lifelong learning opportunities, 2) Self-concept of being an effective digital learner, 3) Initiative, creativity and independent learning concerning digital agriculture learning issues, 4) Self-responsibility in digital agriculture occupations, 5) Optimistic about agriculture's evolution in the digital era, 6) Problem-solving and decision-making concerning agriculture practices in the digital era.

## ACKNOWLEDGMENT

The researchers would like to thank the Faculty of Industrial Education and Technology, King Mongkut's University of Technology Thonburi (KMUTT), and Thailand Cyber University Project (TCU), Ministry of Higher Education, Science, Research, and Innovation for supporting this research.

## REFERENCES

[1] Arvanitis, K. G., & Symeonaki, E. G. (2020). Agriculture 4.0: the role of innovative smart technologies towards sustainable farm management. *The Open Agriculture Journal*, 14(1).

[2] Villalba, M.L. & Elkader, M.A.A. (2020). Logistics 4.0 Technologies in Agriculture Systems: Potential Impacts in the SDG. In Proceedings of the International Association for Management Technology, International Conference on Management of Technology: Towards the Digital World and Industry X.0, Giza, Egypt, 13–17 September 2020.

[3] Bowersox, N. (2011). Encouraging participation in virtual communities of practice within the United States Air Force. In *Social Knowledge: Using Social Media to Know What You Know* (pp. 179-192). IGI Global.

[4] Peeters, W., & Pretorius, M. (2020). Facebook or fail-book: Exploring “community” in a virtual community of practice. *ReCALL*, 32(3), 291.

[5] Wannapiroon, P., Nilsook, P., Kaewrattanapat, N., Wannapiroon, N., & Supa, W. (2021). The Virtual Learning Resource Center for the Digital Manpower. *International Education Studies*, 14(9), 28-43.

[6] Sriprasertpap, K., Langka, W., & Boonlue, S. (2015). The Development of Thai Teacher Information Communication Technology (ICT) Online Community Model in Thailand. *Procedia-Social and Behavioral Sciences*, 197, 1727-1731.

[7] Maneewan, S., Nittayathammakul, V., & Lertyosbordoin, C. (2017, March). A development of knowledge management process on cloud computing to support creative problem solving skill on studio photography for undergraduate students. In 2017 6th International Conference on Industrial Technology and Management (ICITM) (pp. 27-31). IEEE.

[8] Schreier, M. (2012). *Qualitative content analysis in practice*. Sage publications.

[9] Varpio, L., Paradis, E., Uijtdehaage, S., & Young, M. (2020). The distinctions between theory, theoretical framework, and conceptual framework. *Academic Medicine*, 95(7), 989-994.

[10] Yusof, N. I., Zainuddin, N. M. M., Hassan, N. H., Sjarif, N. N. A., Yaacob, S., & Hassan, W. A. W. (2019). A guideline for decision-making on business intelligence and customer relationship management among clinics. *International Journal of Advanced Computer Science and Applications*, 10(8), 498-505.

[11] Materia, V. C., Giare, F., & Klerkx, L. (2015). Increasing knowledge flows between the agricultural research and advisory system in Italy: combining virtual and non-virtual interaction in communities of practice. *The Journal of Agricultural Education and Extension*, 21(3), 203-218.

[12] Anil, B., Tonts, M., & Siddique, K. H. (2015). Strengthening the performance of farming system groups: perspectives from a Communities of Practice framework application. *International Journal of Sustainable Development & World Ecology*, 22(3), 219-230.

[13] Brown, M. E., Ihli, M., Hendrick, O., Delgado-Arias, S., Escobar, V. M., & Griffith, P. (2016). Social network and content analysis of the North American Carbon Program as a scientific community of practice. *Social Networks*, 44, 226-237.

[14] Dolinska, A., & d'Aquino, P. (2016). Farmers as agents in innovation systems. Empowering farmers for innovation through communities of practice. *Agricultural Systems*, 142, 122-130.

[15] Nuutinen, M., & Leal Filho, W. (2018). Online Communities of Practice Empowering Members to Realize Climate-Smart Agriculture in Developing Countries. In *Climate Literacy and Innovations in Climate Change Education* (pp. 67-83). Springer, Cham.

[16] Triste, L., Debryne, L., Vandenabeele, J., Marchand, F., & Lauwers, L. (2018). Communities of practice for knowledge co-creation on sustainable dairy farming: features for value creation for farmers. *Sustainability Science*, 13(5), 1427-1442.

[17] Dolinska, A., Oates, N., Ludi, E., Habtu, S., Rougier, J. E., Sanchez-Reparaz, M., ... & d'Aquino, P. (2020). Engaging farmers in a research project. Lessons learned from implementing the Community of Practice Concept in innovation platforms in irrigated schemes in Tunisia, Mozambique and Ethiopia. *Irrigation and Drainage*, 69, 38-48.

[18] Edwards, A. L., Sellnow, T. L., Sellnow, D. D., Iverson, J., Parrish, A., & Dritz, S. (2021). Communities of practice as purveyors of instructional communication during crises. *Communication Education*, 70(1), 49-70.

[19] Rosmansyah, Y., Achiruzaman, M., & Hardi, A. B. (2019). A 3D multiuser virtual learning environment for online training of agriculture



- surveyors. *Journal of Information Technology Education: Research*, 18, 481-507.
- [20] Auer, T., & Felderer, M. (2020, November). Gamified Internet of Things Testing within a Virtual Learning Environment—towards the Interactive Simulation Game “IoTCityLab”. In *2020 IEEE 32nd Conference on Software Engineering Education and Training (CSEE&T)* (pp. 1-4).
- [21] Lan, P. S., Liu, M. C., & Baranwal, D. (2020). Applying contracts and online communities to promote student self-regulation in English learning at the primary-school level. *Interactive Learning Environments*, 1-12.
- [22] Fenech, R. (2021). Blended learning: Honouring students’ Psychological Contract. *Cogent Education*, 8(1), 1914286.
- [23] Wema, E. F. (2021). Developing information literacy courses for students through virtual learning environments in Tanzania: Prospects and challenges. *IFLA Journal*, 03400352211018231.
- [24] Boulton, C. A., Kent, C., & Williams, H. T. (2018). Virtual learning environment engagement and learning outcomes at a ‘bricks-and-mortar’ university. *Computers & Education*, 126, 129-142.
- [25] Khlaisang, J., & Mingsiritham, K. (2016). Engaging virtual learning environment system to enhance communication and collaboration skills among ASEAN higher education learners. *International Journal of Emerging Technologies in Learning*, 11(4).
- [26] Kingsawat, K., Kwiecien, K., & Tuamsuk, K. (2015). Components and Factors in Integrating Information Literacy Instruction in Elementary Education Using a Virtual Learning Environment. *LIBRES: Library & Information Science Research Electronic Journal*, 25(1).
- [27] Kultawanich, K., Koraneekij, P., & Na-Songkhla, J. (2015). A proposed model of connectivism learning using cloud-based virtual classroom to enhance information literacy and information literacy self-efficacy for undergraduate students. *Procedia-Social and behavioral sciences*, 191, 87-92.
- [28] Sathanarugsawait, B., & Samat, C. (2021, November). Designing of Learning Environment Model to Enhance Student’Self Regulation by Using Massive Open Online Course. In *International Conference on Innovative Technologies and Learning* (pp. 542-547). Springer, Cham.
- [29] Srikan, P., Pimdee, P., Leekitchwatana, P., & Narabin, A. (2021). A Problem-Based Learning (PBL) and Teaching Model using a Cloud-Based Constructivist Learning Environment to Enhance Thai Undergraduate Creative Thinking and Digital Media Skills. *International Journal of Interactive Mobile Technologies*, 15(22).
- [30] Kaeophanuek, S., Na-Songkhla, J., & Nilsook, P. (2019). A learning process model to enhance digital literacy using critical inquiry through digital storytelling (CIDST). *International Journal of Emerging Technologies in Learning*, 14(3).
- [31] Hobbs, R., & Coiro, J. (2019). Design features of a professional development program in digital literacy. *Journal of Adolescent & Adult Literacy*, 62(4), 401-409.
- [32] Pedaste, M., Mäeots, M., Siiman, L. A., De Jong, T., Van Riesen, S. A., Kamp, E. T., ... & Tsourlidaki, E. (2015). Phases of inquiry-based learning: Definitions and the inquiry cycle. *Educational research review*, 14, 47-61.
- [33] Curran, V., Gustafson, D. L., Simmons, K., Lannon, H., Wang, C., Garmsiri, M., ... & Wetsch, L. (2019). Adult learners’ perceptions of self-directed learning and digital technology usage in continuing professional education: An update for the digital age. *Journal of Adult and Continuing Education*, 25(1), 74-93.
- [34] Petcharawises, C. (2012). Development of A Non-formal Education Program Based on Problem-based Learning and Self-directed Learning to Develop Personal Mastery Competency for Head Nurses. *Scholar: Human Sciences*, 4(1).
- [35] Ra, S., Ahmed, M., & Teng, P. S. (2019). Creating high-tech ‘agropreneurs’ through education and skills development. *International Journal of Training Research*, 17(sup1), 41-53.
- [36] Manyuen, M., Boonlue, S., Neanchaleay, J., & Murphy, E. (2019, July). Thai Agriculturists Use of and Experience with Digital Technology. In *2019 The 10th Hatyai National and International Conference* (pp. 1794-1806).
- [37] Pliakoura, A., Beligiannis, G., & Kontogeorgos, A. (2020). Education in agricultural entrepreneurship: training needs and learning practices. *Education+ Training*.
- [38] Bartholomew, S., Strimel, G., Byrd, V., Santana, V., Otto, J., Laureano, Z., & Derome, B. (2020). using data to improve precision in crop fertilization through digital agriculture. *Technology and Engineering Teacher*, 79(7), 32-36.
- [39] Pilotti, M., Anderson, S., Hardy, P., Murphy, P., & Vincent, P. (2017). Factors Related to Cognitive, Emotional, and Behavioral Engagement in the Online Asynchronous Classroom. *International Journal of Teaching and Learning in Higher Education*, 29(1), 145-153.
- [40] Peng, W. (2017). Research on model of student engagement in online learning. *Eurasia Journal of Mathematics, Science and Technology Education*, 13(7), 2869-2882.
- [41] Bond, M., Buntins, K., Bedenlier, S., Zawacki-Richter, O., & Kerres, M. (2020). Mapping research in student engagement and educational technology in higher education: A systematic evidence map. *International journal of educational technology in higher education*, 17(1), 1-30.

# Deep Q-learning Approach based on CNN and XGBoost for Traffic Signal Control

Nada Faqir, Chakir Loqman, Jaouad Boumhidi

Department of Computer Science, Faculty of Sciences Dhar El Mehraz  
Sidi Mohammed Ben Abdellah University of Fes, Fes, Morocco

**Abstract**—Traffic signal control is a way for reducing traffic jams in urban areas, and to optimize the flow of vehicles by minimizing the total waiting times. Several intelligent methods have been proposed to control the traffic signal. However, these methods use a less efficient road features vector, which can lead to suboptimal controls. The objective of this paper is to propose a deep reinforcement learning approach as the hybrid model that combines the convolutional neural network with eXtreme Gradient Boosting to traffic light optimization. We first introduce the deep convolutional neural network architecture for the best features extraction from all available traffic data and then integrated the extracted features into the eXtreme Gradient Boosting model to improve the prediction accuracy. In our approach; cross-validation grid search was used for the hyper-parameters tuning process during the training of the eXtreme Gradient Boosting model, which will attempt to optimize the traffic signal control. Our system is coupled to a microscopic agent-based simulator (Simulation of Urban MObility). Simulation results show that the proposed approach improves significantly the average waiting time when compared to other well-known traffic signal control algorithms.

**Keywords**—Convolutional neural network; extreme gradient; traffic control; traffic optimization; urban mobility

## I. INTRODUCTION

Congestion causes serious social problems such as long distance travel, fuel consumption, and air pollution. Factors that contribute to traffic congestion include the proliferation of vehicles, poor road infrastructure, and poor signal control. But it doesn't stop people from buying cars, and building new road infrastructure is expensive. A relatively simple solution is to improve the efficiency of traffic light control. Approaches to reduce congestion are still priority topics for researchers from different disciplines. There have been several technological developments to help solve these problems; but there is a real need for further study and analysis of the impact of this problem on the daily lives of millions of people.

According to World Bank statistics released in 2017, 64% of the world's oil is consumed in the transportation sector. The sector also contributes 27% of his CO<sub>2</sub> emissions globally. According to the same source, according to 2022 statistics, domestic and international transport already accounts for 20% of global greenhouse gas emissions. It could rise as much as 60% by 2050 [1].

Artificial intelligence plays a very important role in this topic; agent-based simulations have attained a sufficient degree of complexity and scalability and have shown their ability to

manage traffic in the urban environment with other means to improve the quality of life of users.

Some researchers have proposed an optimization method for dynamic routing systems. Although this technique has proven effective in improving computation time, it is rarely used nowadays. Another technique for improving dynamic routing systems is to predict traffic conditions, like in [2] where, a machine learning approach for short-term traffic forecasting was proposed. This approach uses common conditions such as traffic volume, speed, and road segment occupancy to forecast short-term traffic volumes.

Our work presents a contribution to this line of research; and more precisely we will study a simple and always interesting situation: an isolated intersection regulated by traffic lights, which we want to manage by means of an agent capable of exploiting the experience acquired from its learning, possibly representing the admissible traffic conditions. We describe a Convolutional eXtreme Gradient Boosting (ConvXGB) algorithm as a deep reinforcement learning algorithm that automatically extracts all useful functions for adaptive traffic light control from raw real-time traffic data and learns the optimal policy for traffic light control instead of using human-crafted functions. We model the control problem as a reinforcement learning problem. A deep convolutional neural network is then used to extract useful features from the raw real-time traffic data (vehicle positions, velocities, waiting times of each vehicle on the road, traffic light status, etc.) to output the optimal traffic signal control decision. ConvXGB combines the performance of a Convolutional Neural Network (CNN) and eXtreme Gradient Boosting (XGBoost). A key feature of ConvXGB is a systematic strategy for choosing between these two models: XGBoost, widely used by data scientists, is a scalable machine learning system for tree boosting that avoids overfitting. It works well on its own and has proven its prowess in many machine learning competitions. Adding machine learning with CNN, a deep learning class that involves hierarchical learning at multiple different levels, brings clarity.

The simulations are carried out using SUMO (urban mobility simulation tool) of [3] in which we can see the results of the potential regulation of the actions carried out. A considerable reduction in waiting times and vehicle delays is achieved by using the proposed method.

The paper morphology will be as follows: the second section provides a literature research on the applications of Reinforcement Learning (RL) algorithms dedicated to traffic

light control. The third section, we briefly present an explanation of the different techniques used. In the fourth section, we present the studied environment and its concepts. The fifth section describes the structure of the proposed model architecture. The simulation results and performance compared to other reference models are presented in the sixth section and the concluding remarks are given in the last section.

## II. RELATED WORK

In general, traffic signals are based on repeated "cycles", where a cycle is a full rotation of all information provided at the intersection, and consisting of a number of phases. According [4] the "phase" is a group of traffic movements through the intersection; undergoing a "green" time interval, followed by "yellow" time and then "red" time. The dedicated time for each phase follows a fixed plan according to [5] method, or by adopting an adaptive green time scenario for each phase according to the traffic dynamics. The RL method applied to adaptive traffic light control has been proven effective in many papers. Multi-agent systems for solving many traffic management and control problems as in [6] research, a challenge in multi-agent environments is whether to approach traffic control as a collaborative problem/domain. In other words, the local optimum usually competes with the global optimum (the entire road network). Dimensional problems related to learning in multi-agent systems. Each implicitly influences the decisions of other agents. Co-evolution and the role of driver behavior in transport networks. The reference [7] introduced Q-learning algorithm for a single isolated intersection, we know that it is a tabular algorithm dealing with a restricted modeling of the environment based on a finite number of states and actions, which only works for this kind of problem and which will suffer from the curse of dimensionality if we ever think of expanding the state space and giving more details on the traffic state. The reference [8], the authors introduced a RL method with context detection for traffic signal control optimization, the detection of the change of context requires the installation of very sophisticated sensors, which is not at all acquired in various road networks throughout the world. The authors of [8] used a variety of RL algorithms to define the impact of the representation of state and action spaces, as well as the choice of actions and the definition of rewards on traffic models for a real-world intersection, and proved the effectiveness of the phasing variable, especially in large and dynamic traffic models. Most of the works deal with the problem of traffic signal control based on human-designed characteristics such as vehicle queue length and average vehicle delay. Human-designed features are abstractions of raw traffic data that ignore useful traffic information and result in suboptimal signal control. For example, vehicle queue length does not take into account vehicles that are not in the queue but are due to arrive soon. This is also useful information for traffic light control. Average vehicle delay reflects only historical traffic data, not real-time traffic demand.

In this work, we draw on the work of [9] and [10] by establishing a more detailed representation of the state space of the studied environment in order to provide more information of the intersection state. In addition, in this paper we propose a new deep learning model for reinforcement learning problems,

called "ConvXGB" which is a combination of CNN and XGBoost.

The ConvXGB model does not use human features, nor does it use vehicle queue lengths, instead it automatically extracts all useful features from the raw traffic data. Our algorithm works efficiently at realistic intersections. Our ConvXGB consists of several stacked layers of convolutions that learn the properties of the input and can learn the properties automatically. Then XGBoost predicts the class labels in the last layer. The ConvXGB model is simplified by reducing the number of parameters under suitable conditions, as it does not require readjusting the weight values in the backpropagation cycle.

## III. METHODOLOGY

### A. Deep Reinforcement Learning

The treatment of a problem by reinforcement learning (RL) using Deep Neural Networks (DNN) is called deep reinforcement learning (DRL). RL is a goal-oriented machine-learning algorithm aimed at achieving it by interacting with the environment. For a problem handled by RL, the environment is iteratively observed by the agent (in our case the road network), and it takes an action (by changing the phase or the duration of the phase of the signal). As a result, it receives from the environment a reward (in the case of traffic light control: waiting time ...) for the chosen action, the reward will be cumulated with its long-term goal (minimize delay, minimize stops ...). Then according to its rules and the state transition probability, the environment switches to the next state.

The RL agent tries to optimize the correspondence between the state space and the action space called policy, by analyzing the rewards discounted, and accumulated in the long run, and from the application of different action sequences. The agent's policy can be adjusted to converge to an optimal policy, by maximizing the expectation of the long-term reward during the learning process. Fig. 1 shows an RL-agent for traffic signal control.

The value function of the reinforcement-learning problem is the estimate of the reward of each pair of state-action in the long run. The value of environment state is the estimated long-run reward discounted by a factor following the policy, as defined in the Bellman equation as follows:

$$v(s) = \sum P(s', r|s, a)[r + \gamma V_{\pi}(s')] \quad (1)$$

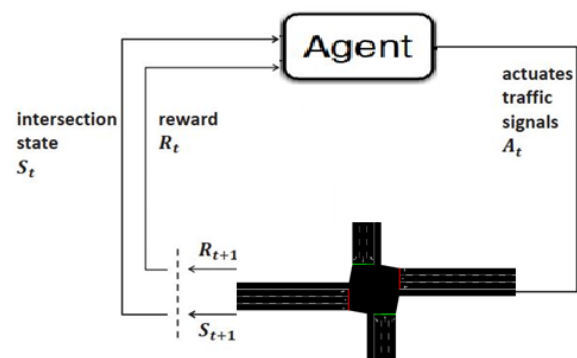


Fig. 1. Reinforcement Learning Agent for Traffic Signal Control.

Where  $s$  represent the environment state,  $a$  the chosen action,  $r$  the reward received from the environment,  $P$  the state transition probability,  $\gamma$  the discount factor,  $\pi$  the agent policy and  $s'$  the next state according the probability  $P$ .

### B. Q-learning

Tabular Q-learning according to [11] is a widely used off-policy RL algorithm. Where the chosen action  $a \in A$  corresponds to the largest Q-value taken from a grid called Q-table. All discrete values of states and actions  $s \in S$  and  $a \in A$  will match in this grid. At each time step, Q-learning improves its policy according to the following equation:

$$Q_t(s, a) = (1 - \alpha)Q_{t-1}(s, a) + \alpha(r + \gamma \max_{a' \in A} Q_{t-1}(s', a')) \quad (2)$$

At time step  $t$  when the agent chooses to execute action  $a$  on states it receives a reward  $r$ , the parameters  $\alpha$  and  $\gamma$  represent respectively the learning rate and the discount factor defining the importance of the upcoming state.

Tabular Q-learning needs a grid to store the Q-values of all pairs  $(s, a)$ . The learning process suffers from the curse of dimensionality using [11], [12] and [13] approaches, if the sets of states  $S$  and actions  $A$  increase and if, in addition, the representation of the state is very detailed, so that the grid of Q values becomes giant. Therefore, the function approximation  $Q(s, a; \theta)$  introduced (where  $\theta$  is the hyper-parameter of the approximator) aimed at the generalization of the function from an example to estimate the correspondence. In this paper, we used this method by convolutional neural networks while adopting a continuous state representation.

### C. Deep Q-Network

The Deep Q-Network (DQN) is a RL algorithm that uses Deep Neural Networks (DNN) as function approximators. Its effectiveness in handling the large state-action space in RL problems as in [14] and in [15] is very significant. In addition, it uses two very important mechanisms to solve instability and divergence problems: experience replay and target network as introduced in [10]. At each learning step, the agent stores the quadruplet  $(s, a, r, s')$  in the replay memory, then it takes a random samples as a mini-batch from the memory to update the weights  $\theta$ . Thus, it eliminates strong correlations between consecutive states. The target network mechanism, which is a neural network identical to the original one, but where weights are updated less frequently, also reduces the correlation problem.

### D. XGBoost

XGBoost (Extreme Gradient Boosting) as explained in [16] is a trendy model widely used in several machine-learning challenges. Indeed, it runs faster than other model and is popular for its scalability in all scenarios discussed in [17]. There are many other boosting algorithms such as parallel boosting, regression tree boosting, stochastic gradient boosting, but in [16], the XGBoost model is one of the leading boosting algorithms.

The performance of many machine-learning algorithms depends on their hyper parameter tuning, it is important to tune a hyper-parameter; we used the cross-validation grid search

technique to tune the hyper-parameters of XGBoost in this paper. Our results show that this step helped our model achieve impressive scores and consequently beat older methods.

### E. Cross-validation Grid Search Tuning Hyper-parameters

XGBoost is a powerful and flexible machine learning algorithm and also it performs very well in general, but there are some problems. Notably the large number of hyper-parameters it has, as well as the fact that different combinations of parameters generate different evaluation scores. Hence it is essential to find the optimal hyper-parameters to get the most out of it. Grid search is method to find optimal hyper-parameters by testing every combination of them. In our case, we choose to tune three common parameters to prevent over fitting: learning rate, minimum child weight and maximum depth.

### F. ConvXGB Model

As a fresh deep learning model for categorization issues, we introduce the Convolutional eXtreme Gradient Boosting (ConvXGB) technique. ConvXGB combines the strengths of a Convolutional Neural Network (CNN) with eXtreme Gradient Boosting (XGBoost) [16], resulting in a state-of-the-art performance and high accuracy, as we will demonstrate. The ConvXGB architecture consists of three sub-CNN-networks each containing three stacked convolutional layers, the outputs of which are merged and reshaped, thus serving as the input to the XGBoost which is the final layer of the model. They differ from conventional CNN in that there is neither a pooling couch nor a fully connected (FC) couch. This adds simplicity and reduces the number of calculation parameters because it is not necessary to adjust the weight in the FC couches in order to adjust the weight in the preceding couches.

## IV. DESCRIPTION OF THE REINFORCEMENT LEARNING ENVIRONMENT FOR TRAFFIC LIGHTS CONTROL

### A. Intersection Model

The environment, on which the agent operates, is a four-armed intersection shown in Fig. 2. Each arm is 500 meters long, the maximum speed is 70km/h and the adopted traffic scenario handles an average of 1000 vehicles per hour. There are three lanes on each arm defining the possible directions of a vehicle: the rightmost lane allows vehicles to turn strictly right, the middle lane allows the driver to go straight only, while on the leftmost lane the driver can turn left or make a U-turn.

In the center of the intersection, the agent controls an adaptive traffic light system. Pedestrians, pavements and pedestrian crossings are not included in our environment.

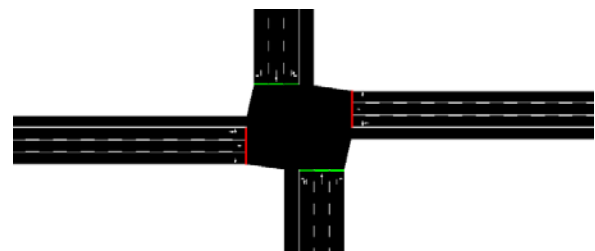


Fig. 2. Intersection Model.

### B. State Representation

A detailed representation of the state of the environment is provided for the RL agent. The state is provided in the form of three matrices on each incoming road: a matrix of the vehicle's position and the matrix of the vehicle's speed used in [10], [18], [19] and [20]; and a waiting time matrix of the vehicle's, and also the last traffic lights status [21]. To create the three matrices  $P_i$ ,  $V_i$  and  $W_i$ , we will need to consider two parameters the segment length  $l$  and the cell length  $c$  for each road  $i = 0, 1, 2, 3$ ; Fig. 3.

In this paper hot coding has been used to define the position matrix of the road  $i$ , the result is the matrix  $P_i$ . The speed matrix  $V_i$  is calculated by normalizing each vehicle's speed by the edge maximum allocated speed. In our case, we consider the waiting time from the moment the vehicle enters the incoming route  $i$  of the network, and the result is recorded at the corresponding entry in the matrix  $W_i$ . The waiting time of a vehicle is defined as the amount of time during which a vehicle has been in the "waiting" state.

From the state of the last activated traffic lights, a matrix  $L$  is generated, where:  $L = [1, 0]$  is a value that defines green lights on horizontal roads whereas  $L = [0, 1]$  represents green lights on vertical roads as used in [10]. In this way, the state representation of the intersection at each time step  $t$  provided to the RL agent will be  $S_t = (P, V, W, L) \in S$ , where  $S$  is the entire state space.

### C. Action Definition

In Fig. 4, at each time step the agent observes the state of the intersection then chooses and executes either action ( $A_t = 0$ ): "0" setting green lights on for horizontal roads or action ( $A_t = 1$ ): "1" setting green lights on for vertical roads. A transition phase is required if the action taken at time step  $t$  is different from the one chosen at time step  $t+1$ , it will be executed as follows:

- 1) Switching the lights for straight ahead vehicles to yellow.
- 2) Switching the lights for straight ahead vehicles to red.
- 3) Switching the lights for left-turning vehicles to yellow.
- 4) Switching the traffic lights for vehicles turning left to red.

The times allocated for the yellow and green lights are fixed and are worth six seconds and ten seconds respectively.

### D. Reward Definition

In reinforcement learning, the feedback that the agent receives from the environment as a result of his choice of action is called a reward. This crucial concept of the training process is used to measure the effectiveness of the choice of action and thus allows the RL agent to improve future choices of his actions.

The reward can generally take positive or negative values. The positive value is the result of the right choice of actions; whereas a negative value is due to the wrong choice of actions. In our case, the objective is to minimize the total waiting times of all vehicles in the intersection. So that the RL agent can measure the effect of the action taken on the effectiveness of

the adaptive control of traffic lights at the intersection in terms of reduction or increase, the reward must be derived from a performance measure of traffic efficiency, so the goal will be met. The intersection is observed for two times, once at the beginning of the time step and then at the end of the time step.

The waiting times in the arrival lanes are noted for each observation and for all vehicles present in the intersection. The formula for the total reward  $R_t$  also used in [10] is therefore:

$$R_t = r_1 - r_2 \tag{3}$$

The values of  $r_1$  and  $r_2$  are respectively the sum of the waiting times recorded at the beginning and at the end of the green light interval according to [10]. The agent will only be rewarded if the value of  $r_2$  decreases; in case a transition phase is mandatory ( $A_t \neq A_{t+1}$ ), the two values  $r_1$  and  $r_2$  will be saved at the end of the execution of the transition phase discussed in section IV.C.

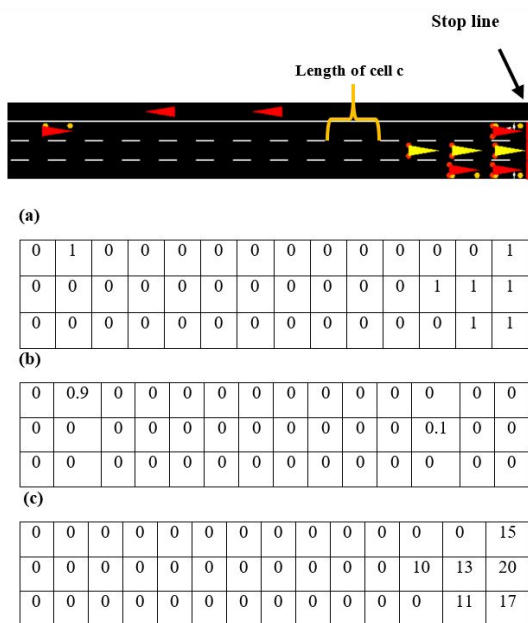


Fig. 3. Illustration of One Arm of the Intersection and the Three Matrices Position (a), Speed (b) and Waiting Time (c) of each Vehicle in the Incoming Lane.

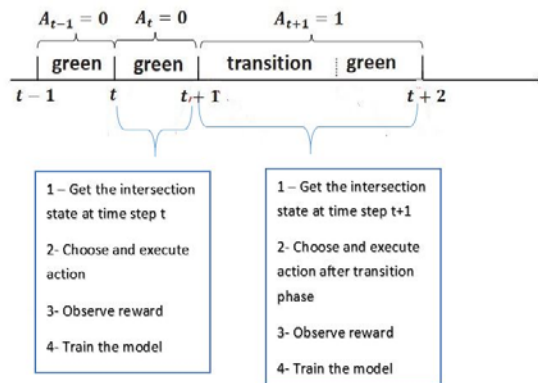


Fig. 4. Chronological Sequence of Events of the Agent: State Observation, Action Choosing, Executing, Getting Rewards and Training Model According.



V. ADAPTATIVE SIGNAL CONTROL ALGORITHM DESIGN  
BASED ON CONVXGB

The ConvXGB agent, which is a RL agent, long-term objective is to reduce congestion at the intersection. After making a series of decisions (actions) in accordance with a policy, we consider  $S_t$ , the intersection's status at the start of time step  $t$ , into consideration. A sequence of rewards  $R_t, R_{t+1}, R_{t+2}, R_{t+3}...$  is given to the agent, after taking a sequence of actions by adopting a policy  $\pi$  for a state  $S_t$  of the environment at the beginning of each time step  $t$ . The action  $A_t$  should maximize the reward  $R_t$  shown in equation (4) to allow the agent to reduce the total waiting times of all vehicles at the intersection at time step  $t$ . The agent is supposed to find an optimal policy for choosing actions, denoted  $\pi^*$  defined in (5) that maximizes the cumulative future reward (Q-values) and that helps it achieve its goal.

$$Q_{\pi}(s, a) = E[\sum_{k=0}^{\infty} \gamma^k R_{t+k} | S_t = s, A_t = a] \quad (4)$$

$$\pi^* \text{ argmax}_{\pi} Q_{\pi}(s, a) \text{ for all } s \in S, a \in A \quad (5)$$

To avoid infinite returns in cyclical processes, rewards must be discounted. Thus at each time step the reward is weighted by a reduction (or discounting) factor  $\gamma \in [0, 1]$ . This discount factor means that the further into the future one is, the less important the rewards become. We denote the optimal values of Q under policy  $\pi^*$  as  $Q^*(s, a) = Q_{\pi^*}(s, a)$ . Our model extract useful traffic features, and predict the Q-values while trying to find the optimal policy  $\pi^*$  and thus find the optimal Q-values  $Q^*(s, a)$ . Bellman's optimality equation (6), gives a recursive relation for the optimal Q-values  $Q^*(s, a)$ .

$$Q^*(s, a) = E[R_t + \gamma Q^*(S_{t+1}, A_{t+1}) | S_t = s, A_t = a] \quad (6)$$

In order to calculate (6), we would need to have complete knowledge of the underlying system mode (state transition probabilities and associated rewards), which is actually not feasible. Due to the complexity and size of the intersection traffic space, ConvXGB model is used to approximatively determine the optimal Q-values. It receives as input the state representation  $S_t = (P, V, W, L)$ . Based on this traffic data, the CNN sub-networks extract useful features, which they then

input to XGBoost to forecast the estimated  $Q(S_t, a)$  for all actions  $a \in A$  given the observed condition  $S_t$ .

A. Proposed ConvXGB Architecture

Our paper is based on a new deep learning model, dedicated to reinforcement learning problems called "ConvXGB". It is a hybridization between convolutional neural network and an XGBoost model. The architecture of ConvXGB is shown in Fig. 5. The model has five layers: 1) input layer, 2) three identical CNN sub-networks each containing three stacked convolutional layers, 3) reshape layer, 4) Q-values prediction layer and 5) output layer. These layers, each of which has unique talents and duties, are crucial to the model's success. Further, the model can be divided into two parts: one for feature learning and the other for Q-values prediction.

These three identical sub-networks having respectively as inputs the matrices (P, V, W). The outputs of these three sub-networks is flattened, and then we merge them with the vector L (last state of the traffic lights) to provide the result to the third layer for reshaping, Table I shows all the CNN parameters.

The second part of the ConvXGB model is an XGBoost model, with tuned parameters using GridSearchCV technique, to predict the Q-values. XGBoost are optimized by a cross-validation grid search.

The performance of the hybrid ConvXGB model is compared to the classical DQN model using only convolutional networks to show its robustness. ConvXGB effectively uses features learning and predicts Q-values, thus significantly reducing the total waiting times for all vehicles in the intersection, which will be detailed in the results section.

TABLE I. PARAMETERS OF THE CNN LAYERS

CNN Layers	Number of filters	Stride	Activation function
Conv1	16 of 4x4	2	ReLU
Conv2	32 of 2x2	1	ReLU
Conv3	64 of 2x2	1	ReLU

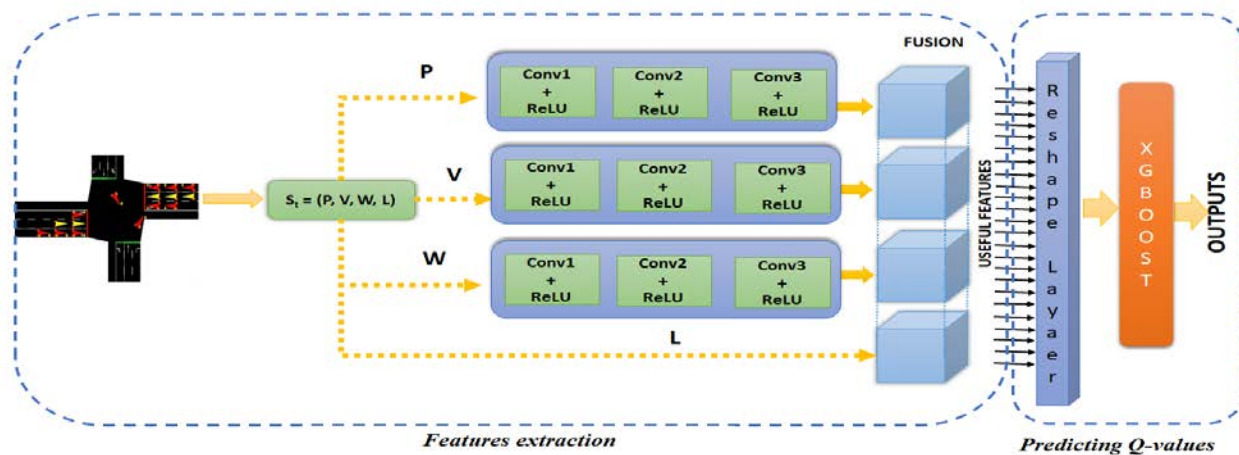


Fig. 5. Architecture of the ConvXGB Model.



## B. Algorithm and Training Process

The algorithm in Fig. 6 summarizes the process of forming the ConvXGB model used. We use a hybrid model composed of three identical CNN sub-networks to extract useful features from an environment representation (state). In the prediction phase, we use XGBoost to make Q-values predictions given the features extracted by CNN sub-networks. We also used the experience replay used in [10] to break correlations and improve agent performance during the training phase. The replay memory provides to the RL agent a set of random experiences for its learning, called a mini-batch. Each experience is represented as a quadruplet  $\{S_t, A_t, R_{t+1}, S_{t+1}\}$ , collected and stored in memory M during the learning phase, after choosing and applying the action  $A_t$  from the state  $S_t$ , the agent receive  $R_{t+1}$  as reward.

The environment undergoes a transition from the state  $S_t$  to the state  $S_{t+1}$ . This is why the observations extracted from this environment are correlated. This correlation can limit the learning capacity of the RL agent and prove the need to use the replay memory. The oldest experiments are erased if the memory is full during learning. The agent needs training data: input data set  $X = \{(S_t, A_t) : t \geq 1\}$  and the corresponding targets  $y = \{Q^*(S_t, A_t) : t \geq 1\}$ . For input data set,  $(S_t, A_t)$  can be retrieved from replay memory M. However, target  $Q^*(S_t, A_t)$  is not known. This formula  $R_t + \gamma \max_{a'} Q(S_{t+1}, a')$  is used to estimate the optimal Q-values. Thus, targets  $y = \{R_t + \gamma \max_{a'} Q(S_{t+1}, a') : t \geq 1\}$ .

The RL agent randomly takes 32 quadruples ( $\{S_t, A_t, R_{t+1}, S_{t+1}\}$ ) of the replay memory used to form training data  $X = \{(S_t, A_t) : t \geq 1\}$  and also to tune and update XGBoost parameters using GridSearchCV technique to efficiently estimate optimal Q-values ( $Q^*$ ).

---

**Algorithm 1** ConvXGB for Adaptative traffic signal control using Q-learning

---

```
Initialize replay memory M
Set the parameters of the CNN sub-networks for learning features
Initialize XGBoost model parameters model with random parameters for the prediction step
Initialize  $\epsilon, \beta, \gamma, episodes$ 
for  $episode = 1$  to  $episodes$  do
  Start new time step
  while  $step \leq maxsteps$  do
    Get current State observation  $S_t$ 
    The CNN sub-networks extract useful features from  $S_t$ 
    The XGBoost model select  $A_t = \text{argmax}_a Q(S_t, a)$  with probability  $1 - \epsilon$  and a random action with probability  $\epsilon$ 
    if  $A_t = A_{t-1}$  then
      No transition phase
    else
      Make a transition phase for traffic signals
    end if
    Execute selected action
    Increment simulation time step
    if time step ends then
      Observe reward  $R_t$  and current state  $S_{t+1}$ 
      Store quadruple  $(S_t, A_t, R_t, S_{t+1})$  into M
      Randomly draw mini-batch of 32 samples from M
      Delete  $M[0]$  if the memory is full
      Form training data using data set X as inputs and y as targets
      Tune XGBoost parameters on training data X and y using GridSearchCV technique
      Update XGBoost model with optimal parameters to predict Q values
    end if
  end while
end for
```

---

Fig. 6. ConvXGB Algorithm for Adaptative Traffic Light Control using Q-Learning.

The exploration / exploitation problem is a frequent problem facing the policy of choice of actions in reinforcement learning; exploration, where we seek more information to improve future decisions; or exploitation, where the decision is made based on current information. In this paper, we have adopted a  $\epsilon$ -greedy exploration policy as used in [18]; given in equation (7). For the current episode  $e$  we have the probability  $\epsilon$  of an exploratory action, and the probability  $1 - \epsilon$  of an exploiting action.

$$\epsilon_e = 1 - \frac{e}{Total\_episodes} \quad (7)$$

At the start of his training, the agent only explores his environment, which is logical and obvious; he begins to use the information received during his training; an exclusive exploitation of the knowledge acquired by the agent takes place towards the end of his training.

## VI. EXPERIMENT AND EVALUATION

### A. Simulation Settings

In this study, we used the SUMO simulator [1], to generate urban traffic simulations. It allowed us to implement and customize road infrastructure functionalities. Moreover, subsequently extract useful data during the traffic simulation.

1) *Intersection*: We have an intersection of four roads, each road having three lanes, as shown in Fig. 2. The length of the road is 500 meters, the maximum speed is 19.44 m/s (i.e. 70 km/h), and the length of the vehicles is 5 meters with a minimum distance between vehicles of 2.5 meters. Moreover, the flow of vehicles (an average of 1000 per hour) is uniformly distributed.

2) *Traffic*: Vehicles, while selecting their route in advance make the choice of entry routes randomly. Horizontal roads will be heavily used while vertical roads, which will be less frequented, we also increase the frequency for left deviating roads compared to right deviating roads. Specially,  $P_1 = 1/7$  (for horizontal roads),  $P_2 = 1/11$  (for vertical roads),  $P_3 = 1/30$  (for roads deviating to the right),  $P_4 = 1/25$  (for roads deviating to the left). The setting of the traffic lights and the transition phase used in this paper are detailed in Section IV.

3) *Agent parameters*: The training process takes place over 2000 episodes. Where each episode corresponds to two hours of traffic and 7000 time steps, simulating the same traffic load at peak hours. For  $\epsilon$ -greedy method in Algorithm 1, parameter  $\epsilon$  is set to be 1 and the discount factor  $\gamma = 0.95$ . Learning rate of RMSProp  $\alpha$  is set to 0.0002 and the replay memory can store the experiences of 200 episodes. The training of the hybrid agent took a few days non-stop high-end laptop.

4) *Simulation data*: The time (in seconds) that a vehicle takes to cross an intersection (between entering and exiting the lanes) defines the delay that it can make. Therefore, the waiting (stopping) time of a vehicle is closely related to its delay. During the simulations, the total waiting time of all vehicles at the intersection is recorded in a file for all episodes.

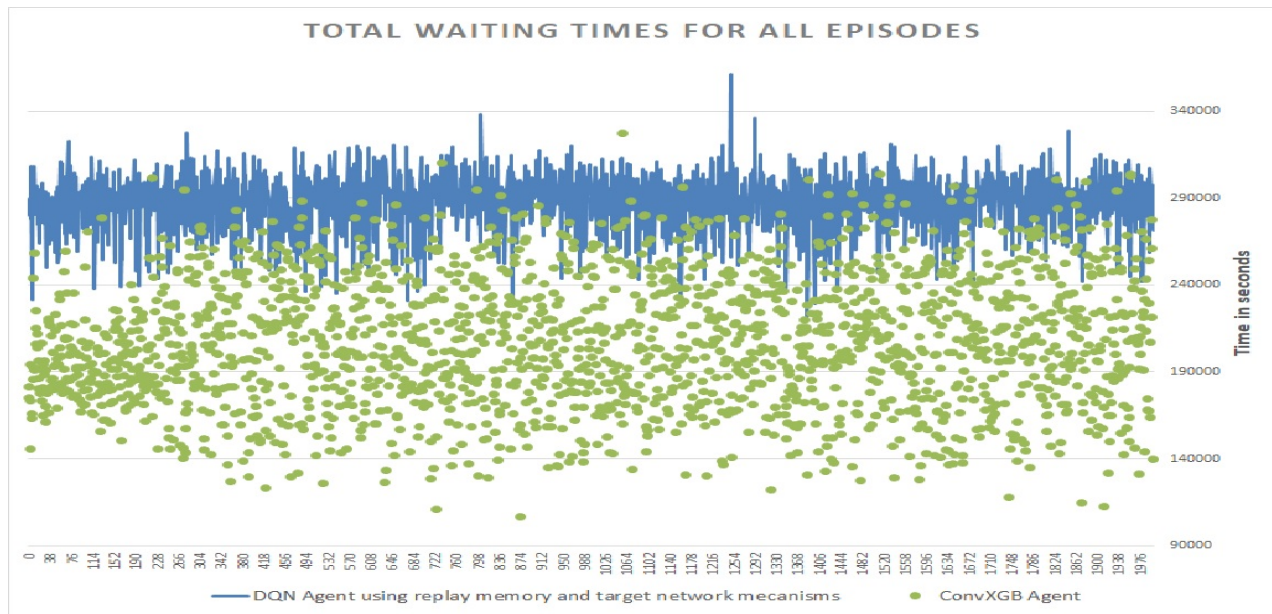


Fig. 7. ConvXGB Agent Improvement.

Our approach is based on a hybrid model as detailed in Section V according to an architecture defined in Fig. 6 and inputs (P, V, W, L) explained in Section IV. The results of our contribution using used a tuned ConvXGB approach, shown in Table II, are compared to the fixed time policy, in which the phase cycle is fixed throughout the day, however the timing of each phase is designed by an expert to accommodate traffic volume ratios and to prioritize arterial traffic for green-waves, and also compared to the approach of [10] that we simulated for our intersection. Approach of [10] showed a clear improvement over the fixed traffic light configuration, but the hybrid ConvXGB agent we developed was able to outperform the fixed time policy and the mentioned approach very well. Fig. 7 presents the performance of the mentioned methods.

Using our agent, we obtain approximately 192k seconds as average waiting times for all episodes, while for the agent in [10] it was approximately 285k seconds. This proves that our ConvXGB hybrid model with XGBoost hyper-parameter tuning has significantly improved the problem of reducing traffic congestion.

TABLE II. TOTAL WAITING TIMES OF ALL VEHICLES FOR THE IMPLEMENTED APPROACHES

	Mean of total WT of all vehicles in seconds	Best total WT of all vehicles in seconds	Improvement over the static scenario
DQN agent [10]	285503	231444	54%
ConvXGB agent	<b>192213</b>	<b>145425</b>	<b>72%</b>
Fixed time policy	505850		

B. Training Evaluation Results and Discussion

The problem of this study was as follows: can a hybrid model "Convolutional Neural Network-Extreme Gradient

Boosting" statistically outperform the "DQN" model in optimizing the control of traffic lights in the urban environment?

The quantitative and complete description of the traffic situation provided to the agent helped a lot. It is clear that the states are more complex but the agent gains in performance. The inputs of our model were designed in a suitable way, in contrast to [22] where the authors used binary position matrices. They defined the binary matrix to cover the entire rectangular area around the intersection instead of just covering the area of the street relevant to traffic light control. Most entries in the binary matrix are null and redundant, making the binary matrix inefficient because the vehicle can't travel on non-road areas. In our case the vehicle position matrix covers only intersecting roads. This reduces the cost of training computation. We were able to boost the performance of our ConvXGB model by using the GridSearchCV technique for fine-tuning the XGBoost model. And since the tests we were able to do this step really helped our model to show better results.

Based on the evaluation and comparison of the two models, the results clearly showed the performance difference of the ConvXGB hybrid model. As shown in Fig. 7, the results obtained are significantly larger than the approach of [10]. The ConvXGB-based Q-learning algorithm is an effective choice over traditional traffic control methods, solving the problem of traffic congestion in large cities.

In the future, a better prediction model could be developed by using a heuristics methods for fine-tuning of the XGBoost and experimenting with new hybrid algorithms is recommended for future work.

VII. CONCLUSIONS

We have developed a new deep learning model for traffic light control problems using reinforcement learning. ConvXGB has two parts: one for useful feature extraction based on

detailed traffic situation definition and one for predicting Q values according to Q learning algorithm. We evaluated ConvXGB, on a single intersection, by adopting a traffic scenario illustrating clear peak hour congestion. ConvXGB was simplified by reducing the number of parameters needed and did not require back-propagation in the fully connected layer. ConvXGB based on CNN and XGBoost, was improved by tuning the most common XGBoost hyper-parameters. Our experimental results show that our model is significantly better than the classic DQN model, which also performed well when comparing it to the fixed time policy, but its performance is clearly inferior to our ConvXGB.

#### REFERENCES

- [1] "Transport overview." [Online]. Available: <http://www.worldbank.org/en/topic/transport/overview#1>. [Accessed: 23-sep-2022].
- [2] O. Mohammed and J. Kianfar, "A Machine Learning Approach to Short-Term Traffic Flow Prediction: A Case Study of Interstate 64 in Missouri," *2018 IEEE Int. Smart Cities Conf. ISC2 2018*, pp. 1–7, 2019, doi: 10.1109/ISC2.2018.8656924.
- [3] M. Behrisch, L. Bieker, J. Erdmann, and D. Krajzewicz, "SUMO—simulation of urban mobility: an overview," *Proc. SIMUL 2011, Third Int. Conf. Adv. Syst. Simul.*, 2011.
- [4] S. Touhbi et al., "Adaptive Traffic Signal Control: Exploring Reward Definition for Reinforcement Learning," *Procedia Comput. Sci.*, vol. 109, pp. 513–520, 2017, doi: 10.1016/j.procs.2017.05.327.
- [5] Webster, "Traffic Signals Webster," *Road Res. Tech. Pap.*, no. 39, pp. 1–44, 1958.
- [6] A. L. C. Bazzan, "Opportunities for multiagent systems and multiagent reinforcement learning in traffic control," *Auton. Agent. Multi. Agent. Syst.*, vol. 18, no. 3, pp. 342–375, 2009, doi: 10.1007/s10458-008-9062-9.
- [7] B. Abdulhai and L. Kattan, "Reinforcement learning: Introduction to theory and potential for transport applications," *Can. J. Civ. Eng.*, vol. 30, no. 6, pp. 981–991, 2003, doi: 10.1139/103-014.
- [8] D. De Oliveira et al., "Reinforcement learning-based control of traffic lights in non-stationary environments: A case study in a microscopic simulator," *CEUR Workshop Proc.*, vol. 223, 2006.
- [9] S. El-Tantawy and B. Abdulhai, "Comprehensive analysis of reinforcement learning methods and parameters for adaptive traffic signal control," *Proc. Transportation Research Board 90th Annual Meeting, Washington DC, USA, Jan. 23-27-2011*.
- [10] J. Gao, Y. Shen, J. Liu, M. Ito, and N. Shiratori, "Adaptive Traffic Signal Control: Deep Reinforcement Learning Algorithm with Experience Replay and Target Network," pp. 1–10, 2017.
- [11] C. Watkins and P. Dayan, "Q-learning," *Machine learning*, vol.8, pp. 279–292, 1992, doi:10.1007/BF00992698.
- [12] C. El Hatri and J. Boumhidi, "Traffic management model for vehicle re-routing and traffic light control based on multi-objective particle swarm optimization," *Intelligent Decision Technologies*, vol.11, no.2, pp. 99–208, 2017, doi:10.3233/IDT-170288.
- [13] M. Tahifa, J. Boumhidi, and A. Yahyaouy, "Multi-agent reinforcement learning-based approach for controlling signals through adaptation," *Int. J. Auton. Adapt. Commun. Syst.*, vol. 11, no. 2, pp. 129–143, 2018, doi: 10.1504/IJAACS.2018.092019.
- [14] V. Mnih et al., "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015, doi: 10.1038/nature14236.
- [15] S. S. Mousavi, M. Schukat, and E. Howley, "Deep Reinforcement Learning: An Overview," *Lect. Notes Networks Syst.*, vol. 16, pp. 426–440, 2018, doi: 10.1007/978-3-319-56991-8\_32.
- [16] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, vol. 13-17-Aug, pp. 785–794, 2016, doi: 10.1145/2939672.2939785.
- [17] C. Bentéjac, A. Csörgö and G. Martínez-Muñoz, "A comparative analysis of gradient boosting algorithms," *Artificial Intelligence Review*, vol. 54, pp. 1937–1967, 2021, doi: 10.1007/s10462-020-09896-5.
- [18] W. Genders and S. Razavi, "Evaluating reinforcement learning state representations for adaptive traffic signal control," *Procedia Comput. Sci.*, vol. 130, no. July, pp. 26–33, 2018, doi: 10.1016/j.procs.2018.04.008.
- [19] W. Genders and S. Razavi, "Using a Deep Reinforcement Learning Agent for Traffic Signal Control," *ArXiv*, vol. abs/1611.01142, pp. 1–9, 2016.
- [20] K. Shingate, J. Komal and D. Yohann, "Adaptive Traffic Control System using Reinforcement Learning," *Int. J. Eng. Res.*, vol. V9, no. 02, pp. 443–447, 2020, doi: 10.17577/ijertv9is020159.
- [21] H. Wei, G. Zheng, V. Gayah, and Z. Li, "A Survey on Traffic Signal Control Methods," *arXiv learning*, vol. 1, no. 1, 2019.
- [22] E. Van Der Pol, "Deep Reinforcement Learning for Coordination in Traffic Light Control," *Master thesis*, no. November 2015, p. 2016, 2016.

# Automatic Text Summarization using Document Clustering Named Entity Recognition

Senthamizh Selvan.R, Dr. K. Arutchelvan

Assistant Professor, Department of Computer and Information Science  
Annamalai University, Chidambaram, Tamil Nadu, India

**Abstract**—Due to the rapid development of internet technology, social media and popular research article databases have generated many open text information. This large amount of textual information leads to 'Big Data'. Textual information can be recorded repeatedly about an event or topic on different websites. Text summarization (TS) is an emerging research field that helps to produce summary from a single or multiple documents. The redundant information in the documents is difficult, hence part or all of the sentences may be omitted without changing the gist of the document. TS can be organized as an exposition to collect accents from its special position, rather than being semantic in nature. Non-ASCII characters and pronunciation, including tokenizing and lemmatization are involved in generating a summary. This research work has proposed an Entity Aware Text Summarization using Document Clustering (EASDC) technique to extract summary from multi-documents. Named Entity Recognition (NER) has a vital part in the proposed work. The topics and key terms are identified using the NER technique. Extracted entities are ranked with Zipf's law and sentence clusters are formed using k-means clustering. Cosine similarity-based technique is used to eliminate the similar sentences from multi-documents and produce unique summary. The proposed EASDC technique is evaluated using CNN dataset and it shown an improvement of 1.6 percentage when compared with the baseline methods of Textrank and Lexrank.

**Keywords**—Named entity recognition; text summarization; k-means clustering; Zipf's law

## I. INTRODUCTION

Internet technology paves the way for information resources. Excessive data is in text form and contains a lot of hidden information. Natural Language Processing (NLP) is one of the basic techniques used to extract hidden information from large amounts of textual data. Social media such as news channels and Facebook play an important role in generating high level text information. Reading huge amount of information is difficult task for human. Text summarization (TS) is the interesting research field that helps to generate summary from the text documents. Due to the data deluge and public consumption, information in the social media has redundant information in all normal dialects. For generating the summary, it is sometimes conceivable to delete words, expressions, rules and complete sentences without disrupting the significance of the message.

TS is the process of creating only brief outlines of the text without redundancies. Statistical analysis has its own limitations with the use of traditional ontological methods for deriving summary [1]. The content writers have different

perspective and use their own writing styles. Hence, traditional ontology does not support to find the cognisance of the knowledge in the documents. This makes more complex for extracting the summary from the documents. Also, the text summary cycle may require an alternative overview of the information. Homogeneous depiction empowers a single representation to gather information from different form of resources using primary integration. This will help in summary productivity. In general, text summarization process is based on finite state automation [2].

These days, a huge number of research articles, news articles, blogs and forums are distributed in every field of study. It represents a great challengeable task for industry experts and researchers to know the latest developments in their specific fields. A new report reveals that many logical articles are copied at regular intervals [3]. The solution of logical essays overcomes this challenge by providing significant findings and commitments to the essay. In the scientific document summary generation, the summary generated by experts have help them to reduce the data collection (secondary research articles) work. It also reduces the work and time required to review any logical article. This was the essential inspiration to run this work.

In general, the text summarization can be done in two ways to get a summary of single or multiple documents. In the primary way, the paper's theory is considered an outline, although the problem with the theory is that it does not reveal the immeasurable significant commitments and findings of an article as a result of length limitations. The findings and confirmations made by the author of an article may be essential from the writer's point of view, but it may not be relevant to the local area. In addition, there is no data from all parts of the theoretical article [4]. The next method overcomes the weaknesses of the main system. Given a note sheet (sheet to be compressed), a note-based summary is generated using the notes in the note sheet. This approach has negative pages based on various authors composing reference texts, and any misconception by them may present mistakes in the last outline; As a result, facts, findings, and basic structure of the reference sheet may be missed. Reference Ecology [5] can take care of this problem. Here every sentence referring to the note-taking notes will be removed first. This arrangement of sentences is called reference systems. It can create the last summary using extracted sentences.

As it is discussed above, two methods are used to obtain the outline of the text: extractive and abstractive [6]. The Extractive Text Summarization (ETS) technique combines the

deletions obtained from the corpus to the frame outline. The Abstractive Text Summarization (ATS) technique creates new sentences from the data obtained from the corpus. The ETS technique is inapt for multiple document resolution. The cause is the likelihood of creating a one-page summary of a few sources [7]. Again, little effort has been made to summarize the evaluation records and differentiate between the components that affect the presentation of each method. Valuation logs contain ratings and preferences, e.g., websites or client surveys.

The goals of this research work are as follows:

- Proposing an entity ranking method to rank the important entities.
- To find the sentence similarity using Cosine similarity.
- Grouping the sentences of multi-document using k-means clustering.
- Generate the text summary using entity aware and document clustering.

The organization of this research article is as follows: the detailed overview of text summarization and its methods are discussed in this introduction Section I. The literature review of previous work on text summarization on various applications such as clinical summary generation, news articles summary generation and scientific article summary generation are discussed in the Section II. The proposed EADSC method is given with a big picture in the Section III. Results of the proposed work is discussed in the Section IV. Finally, this research article is concluded with the future work in section V.

## II. BACKGROUND STUDY

The introduction of the Transformer encoder-decoder models briefly sparked [8], [9] news [10] and logical articles [11] and significant improvements. By the way, their application for summary of medical notes has not been satisfactorily examined. A prototype in the light of Pointer-Generator-Networks [10], [12] has been proposed for a concise outline of radioactivity by combining materials in the medical specifications of UMLS [13] and RadLex [14]. They use inventions and impression pairs for abstract work, where inventions form communications and create objective definitions for creating records.

Sotudeh et al. [15] have proposed a two-level model that includes material selection and abstract summary for medical abstraction. Selector is ready to distinguish ontological terms from discoveries through medical metaphysics (Radlex) and create concise records. Two-LSDMs are used to encrypt inventions and LSDMs are used to create solutions following the LSDM-based decoder. Liang et. al [16] have developed a model for differentiating clinical symptoms in patients with diabetes and hypertension and developing obvious contractions of the disease. They examined a database of 3,453 medical records collected for 762 patients, outlining the difficulty in determining age as a punishment. The authors [17] have proposed a model that incorporated the syntax-based misdiagnosis and approval of the syntax medical idea for extracting the medical message. They conducted their experiments on the MIMIC-III [18] database.

Text report outline is the focus of much research in the research field of NLP. Different kind of methods have been used to solve this problem, such as passive semantic investigation [19], object visualization and poison models [20] and the meta-heuristic method [21]. The essence of live models is the control of information, because a lot of coded information is needed to solve a brief task [22]. In recent days, many researchers involved in developing graph-based models for generating summary from multi-documents. Sentences are considered as vertex or nodes, and the margins between the vertex indicate the similarity between the sentences. The key aspect of the graph-based approach is to determine the most focal sentence in the record. Part of the graph-based models are LexRank [23] and TextRank [24]. The logical solution from the bat made by Duffel et al is correct. [25].

The benefit of using the reference method to create the exterior of a research paper has been demonstrated by Elkis et.al. [26] and Hernandez et.al. [27]. Hong et.al. [28] have developed logical outline practices using various stabilized material wood. Cohen et al. [29] proposed a search-organized approach to separating significant sections of the reference sheet. Most experts are involved in reference contextualization to create a logical solution to the dissertation. The shared tasks in the TAC 2014 database, CL-SciSumm 2016 and the 2017 Logical Report solution [30] provided databases for local education for select reason. Although the CL-SciSumm 2016 and 2017 datasets compile the logical articles of the Computational Semantics section, the TAC database is linked to the Biomedical Article solution.

In a research work, the authors [31] have developed a group-based methods. The important notes are compiled first and more focused phrases are selected from different clusters to create the exterior. Li et al. [32] SVM classifiers are used to outline vocabulary and sentence proximity. The authors [33] have published a semi-ethnic model using similarities based on brain structure and tfidf, which scores relevant text that can be memorized for abstraction. The developers [34] proposed a revised TextRank calculation called TextSentenceRank to sort the sentences; here, a solution is designed in view of the stabilized sentences.

Baki et al. [35] Cosine similitude was used using the term frequency-inverse document frequency vector and a subgroup using SVM (Support vector Machine) [36]. The SVM is combined with Decision Tree (DT) to identify the reference length. The outline is created by removing the sentences considering the average notch in each quote. In another research work, the authors [37] have used TF as a vector space model that uses a characteristic approach to the representation of text and a non-negative structure. Lapalme G et al. [38], used similitude ability to identify reference text, including tips and very serious small fit to create the solution. Cao et al. [39] was used SVM model to rank the reference text for each reference, and the final solution was prepared by complex stabilization [40]. The authors [41] have worked on the tfidf comparison, jacquard similitude and proximity system were used to differentiate the reference length, which is further applied to the outline.



Chiruzzo et al. [42] have experimented the ACL Collection Note that uses a variety of embeds, such as Corpus Word Matching [43] and Google Newsword Installation. Jacquard used similarity [44], LDA [45] and more, with less emphasis on reference identity intimacy and outline age. Glavas et al. [46] the design of the reference summary used element-based features, level-based content, vector space likeness and unigram to obtain text-range. Dipankar et al. [47]. The cosine proximity was used to differentiate the reference length for the solution from the point of view of the scoring system.

The author proposed a work that, [48] reference length was extracted using a standard language model [49], printed content [50] and basic writing learning [51], which are additionally used to relate to the abstract design score. Cohen et al. [52] proposed a logical abstract technique. In another work, similar authors [5] proposed a technique for applying the reference context using question correction, word formation and direct learning. In another work, the researchers [53], have a separate reference structure using cosine likeness and jacquard comparison, and selects phrases in the light of different highlights from the note sheet to create a summary that is closely linked to the notes.

In another research work, authors have proposed a framework for logical resolution [54] to recognize the reference system using the mover's distance and the LDA model; also, they used a set age process (DPP) for the abbreviated age. The authors have proposed a framework [55] and used weighted democratic classifiers to extract the citation system. Clustering method was used to generate the summary. The researchers [56] have used the open NMT an application for the TS. Nghiem et al. [57] an adjusted two-way transformer was used to differentiate the reference system. Furthermore, they have proposed a semi-ethnic outline technique for the compression age. The table I represents the scope and drawbacks of the different automatic text summarization approaches.

TABLE I. SCOPE AND RESTRICTIONS OF THE CONVENTIONAL ATS APPROACHES

Reference	Year	Description	Limitations
[58]	2009	important ways to summarise the content and a taxonomy of summarising techniques	Missing from NLP is extractive, abstract, machine learning, and deep learning.
[59]	2014	Reviewed works from 2000 to 2013 and suggested a statistical approach.	excludes cognitive components
[60]	2014	A hybrid strategy can effectively use both extractive and abstractive Techniques	avoided difficult procedures
[61]	2016	Introduces the concepts of abstractive and extractive summarization.	only describe strategies and procedures
[62]	2017	Extractive approaches for summarising multilingual texts are presented.	There is a lack of a definite classification and concept of feature score.
[63]	2021	to manage several materials for comparison and summary based on recent research work	does not constitute a quick conversation

Li et al. [64] have distinguished between the Word 2 reference system for the CNN model and the determining point processes (DPP) for the text summarization. Cagliero et al. [65] have used a writing model for the reference system for individual sources. The short path amongst the selected text length is predicted and the summary is generated. The authors [66] have used a variety of classification and ballot systems to identify the reference system. To format the summary, they compile phrases and select high-quality phrases from each encounter. Researchers [67] have diagnosis of the use of intermediate brain systems and how to monitor the reference system; for a long time, abstraction was designed by selecting phrases such as notes.

### III. MATERIALS AND METHODS

Extracting a summary from multi-documents are created by cutting key pieces of text from the collected documents. Statistical analysis is involved to find the important of sentences. The overall process involves in the proposed EASDC technique is given in the Fig. 1.

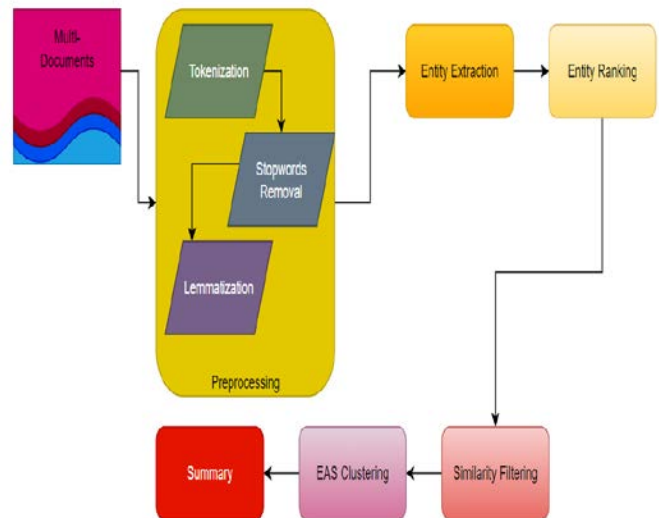


Fig. 1. Research Process of Proposed EASDC Work.

The proposed EASDC technique works using the following components: pre-processing, entity extraction, entity ranking, similarity filtering and EAS clustering. These components are well explained in the following sub-sections.

#### A. Pre-processing

In the text mining, preprocessing plays a major role to provide the documents in structured representation. Multi-document is defined by the following equation (1).

$$\left\{ \begin{array}{l} \{d_1, d_2, d_3, \dots, d_n\} \in MD \\ \text{if } d > 1, MD \text{ is true} \\ \text{if } d > 0 \text{ and } d \leq 1, MD \text{ is false} \end{array} \right\} \quad (1)$$

From the equation 1, d resembles document and MD resembles collection of documents. If d is lesser and equal to one, then it is single document. In English language, each sentence is identified by the full stop (.) at the end of a word. In preprocessing step, each sentence is tokenized using the regular



expression. The sentence tokenization is given in the equation (2), if number of sentences is lesser than five then the summarization process is not needed.

$$\left\{ \begin{array}{l} \{sen_1, sen_2, sen_3, \dots, sen_n\} \in S \\ \text{if } sen \geq 5, S \text{ is true} \\ \text{else, } S \text{ is false} \end{array} \right\} \quad (2)$$

The preprocessing phase involves the following technical process:

1) *Tokenization*: Each word is classified by a token in the sentence. As it is given in the equation 2, each sentence is classified by a token in a document. The sentences are tokenized and store into a desired format. In this proposed work, the tokenized sentences are stored in the array format and resembled as a sequence of tokenized-sentences. Removing non-ASCII characters are essential before proceeding the tokenization process. The non-ASCII characters are meaningless and it is not necessary for the text mining process.

2) *Removal of stopwords*: After tokenization, stopwords are important to remove the useless word in the sentences. For generating the summary, the subjective words are sufficient to find the important sentences. Hence the useless words such as the, of, a, an, in can be removed and generate new tokens.

### B. Entity Extraction

Entities are the key terms that helps to identify the important sentences. The word ambiguity is the important challenge in the entity extraction. Named Entity Recognition (NER) is the important task in NLP to extract the important keywords. For example, consider the following sentence, Chennai super kings won the T20 match that was held in Chennai. In the given sentence, Chennai is location and Chennai super kings is an organization. The output of the entity extraction must be as follows: Chennai super kings (Organization) won the T20 match that was held in Chennai (location). The ambiguous words can be identified by the human easily, whereas the human generated computational application needs lot of training.

For this research work, the fastest Spacy 3.0 library is used to extract the entities. The extracted entities are converted into numerical using token-2-vector model that is present in the spacy library. Conditional Random Field (CRF) technique is used to tag the ambiguous word with its identified entities. The CRF can be calculated using the following equation (3).

$$p(y|X) = \frac{1}{Z(X)} \prod_{n=1}^N \exp\{\sum_{m=1}^M \delta_m f_m(y_n, y_{n-1}, X_n)\} \quad (3)$$

where y is the part-of-speech of the current token and X is the observed entity. The weight of the words are resembles using  $\delta_m f_m$  and  $y_n, y_{(n-1)}, X_n$  resembles the features of the sentences. Z(X) is the total quantity of the named entities (NE). The weight estimation of the NE token is calculated using the maximum-likelihood estimation.

### C. Entity Ranking

NE ranking is used to identify the importance of a NE present in the documents. In general, the frequency of NE across the document gives the importance of a NE (i.e., number of appearances of NE in the document). The actual frequency of word has to be calculated using the inverse proportional of entity in the whole document. Zipf's law is one of the popular methods to identify the rank of the word in the given document. Zipf's law states that, the rank frequency of word is inversely proportional to the rank in the frequency table. The Zipf's law is defined in the equation (4) given below,

$$Z(r, \beta) \propto \frac{1}{r^\beta} \quad (4)$$

where  $\beta \approx 1$ , r denotes the rank of word and Z(r,β) represents the frequency of entity in documents. The identified entities are ranked and organized using the above equation 4.

### D. Similarity Filtering

Cosine similarity is calculated with the following equation (5) given below. It is used to find similarity in the sentences. From the equation (4), if the rank value is lesser than 0 is not considered for the document summary and it is eliminated from the tokenized entities. For the proposed EASDC work, the threshold value for the similarity index is set to greater than 75 percentage to get perfect similarity [68].

In the equation (5), sentence similarity is evaluated by comparing a sentence with all other sentences. The similarity matrix is created based on the values. The highly identical sentences based on the cosine value are removed from the documents. This reduced the redundant information from the multiple documents.

$$\cos(s_i, s_j) = \frac{\sum_{k=1}^n s_i * s_j}{\sqrt{\sum_{k=1}^n s_i^2} \sqrt{\sum_{k=1}^n s_j^2}} \quad (5)$$

where S is sentences and si and sj denotes the current and next sentence respectively.

### E. EAS Clustering

Sentences in multiple documents are ranked based on the presence of entities. The entities are ranked by using the equation 4. The duplicate sentence is eliminated using the equation 5. The unique sentences are embedded using the Doc2Vec mechanism. It is the extension of word embedding. For sentence embedding, this research work uses the Distributed Memory version of Paragraph Vector (PVDm) technique. The embedded sentence is represented with a unique token and stored in a matrix format.

The document clustering is one of the popular methods that is used to group the documents. In the proposed EASDC research work, it used k-means clustering mechanism to grouping the sentences based on the cosine similarity. Cosine distance method is applied to find the distance between the similar sentences. The equation for the cosine distance is given in the equation 6.

$$\cosine_{distance} = \frac{\sum_{n=0}^N s_n - p_x}{\sum_{n=0}^N (s_n)^2 * \sum_{n=0}^N (p_x)^2} \quad (6)$$

where  $s_n$  is the sentence with the position  $n$ , where  $n$  is 1, 2, 3...,  $N$ .  $N$  is the total number of sentences.  $p_x$  is the next sentence with point of observed  $x$ . The pseudocode of  $k$ -means-clustering pseudocode is given below.

---

**Pseudocode 1:**  $k$ -means clustering

---

**Input:** embedded sentences  
**Output:** grouped sentences  
Set number of  $k$   
Set centroids  $p_1, p_2, \dots, p_x$  randomly  
Repeat steps 4 and 5 till the end of iterations  
for  $n$  in  $p_x$ :  
find the nearest centroid  
assign the point to that cluster  
for  $j$  in cluster\_ $k$ :  
find new centroid by calculating the mean of centroids  
End

---

The clusters are pre-defined based on the length of summary that has to be created. The vector of the sentences is used to find the distance between the sentences. The number of clusters is set based on the entities and the chosen topic.

The clustered sentences may have different topics, because, the cluster is formed based on the entities and topics that has discussed in the documents. Therefore, each cluster may have a higher probability of being different topics. The top sentence in each cluster is considered and helps to form the summary. Each sentence from a cluster is sufficient to exaggerate the important of the topic. If there is one cluster, then the sentences in the cluster is extracted to form the summary. In another situation, if similarity is found between the topics, the top sentence of each cluster is taken and the similarity of the sentences is calculated. The redundant sentence from those clusters is eliminated and summary can be generated. The summary generation is given in the pseudocode below.

---

**Pseudocode 2:** Summary Generator

---

**Input:** clustered sentences  
**Output:** Summary  
Load the clustered sentences as  $C$   
Set  $T\_C = [ ]$  //list of topics, i.e., cluster head.  
Set  $summary = [ ]$   
For  $j$  in  $C$ :  
If  $[T\_C(j)]$  and  $[T\_C(j+1)] > 0.75$  : #  
 $T\_C(j)$  denotes,  $j$ th sentence of a cluster  $C$  under topic  $T$   
For  $s$  in  $j$ :  
Summary.append( $s$ )  
Break  
If  $[T\_C(j)]$  and  $[T\_C(j+1)] < 0.75$  :  
Eliminate the redundant sentence  $T\_C(j)$   
End

---

#### IV. RESULTS AND DISCUSSIONS

The proposed EASDC method generates summary using document clustering and named entity recognition. The identified entities are ranked and create the cluster using the equation 5 and 6. The clustered sentences are involved in generating the summary. The summary of the multi-documents

is restricted based on the required length. The sentence similarity plays major role in eliminating the redundant sentences. The sentences are embedded and it helps to diminish computational sparsity.

To evaluate the recital of the proposed EASDC research work CNN dataset is used. The entity extraction played a major role in the proposed research process. It turned out to be more difficult than we had anticipated. The DUC dataset looks to have a perfectly functioning XML structure, but we were unable to load it using numerous Python modules because of errors in the XML format. After that, we had to separate the lengthy text dumps into sentences. Our initial, basic implementation simply divided the text into paragraphs using periods. To speed up the loading of data, this data was saved to disc. This process can take over an hour to perform from scratch, but once finished, it doesn't need to be repeated.

The entities are extracted using the python spacy library. The named entities such as person, organization, location, GPE (geographical place), date, time and money. These entities are ranked based on their occurrences using Zipf's laws. The document-clustering using  $k$ -means cluster helps to grouping the sentences and generate the summary.

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) is used to measure the efficient of proposed summarization technique, which is given in the equation 7.

$$ROUGE - N(c, r) = \frac{\sum_{r_i \in r} \sum_{n-gram} r_i Count(n-gram, c)}{\sum_{r_i \in r} numNgrams(r_i)} \quad (7)$$

Where  $N$  is the grams or tokens or words,  $c$  and  $r$  is candidate and reference respectively.

In this article, the following metrics are tested to evaluate the efficiency of the proposed EASDC technique:

- ROUGE - 1
- ROUGE - 2
- ROUGE - L

Table II denotes the outcome of comparison between EASDC with TextRank and LexRank methods. The proposed EASDC method outperformed well.

The comparative outcomes of the tested strategies for graphical representation are shown in Fig. 2. The LexRank algorithm yielded an average of 38.3%, compared to 39.06% for the TextRank algorithm. The proposed EASDC method performed better and generated 40.73%. It demonstrated a 1.67 percent improvement over the TextRank algorithm. The ROUGE scores improved with entity extraction-based extractive summarization.

TABLE II. SCOPE AND RESTRICTIONS OF THE CONVENTIONAL ATS APPROACHES

	LexRank	TextRank	EASDC
Rouge - 1	36.6	37.6	39.2
Rouge - 2	37.4	38.4	39.9
Rouge - L	40.9	41.2	43.1

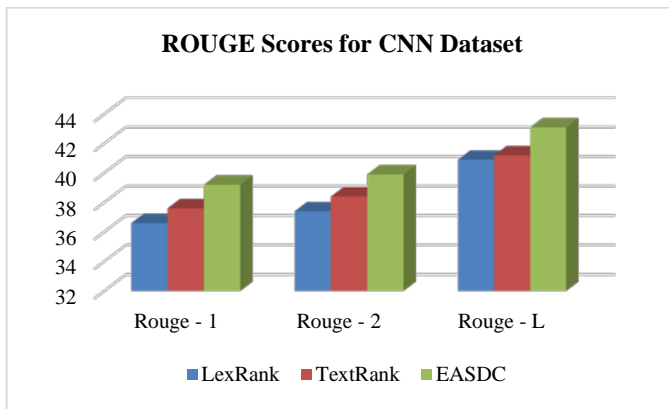


Fig. 2. ROUGE Scores of Proposed EASDC Technique and Existing Works.

## V. CONCLUSION

The size-sensitive expansion of the World Wide Web has created better access to textual information. This work presents a technique for generating literary information that solves the problem of repetition and error in one of these ways. Implementing the proposed framework is not significantly different from business text summaries. The entity ranking and sentence similarity calculation helps to extract the unique sentences from the multiple documents. The extracted NE are then passed to the document clustering methods. By estimation, k-implies are a group calculation and high-level cluster calculations are used for incomparable effects. Similarly, the tendency to extract the sentence from each group is not based on random correlation rather to develop a particular calculation. The proposed EASDC technique shown an improvement of 1.67 percentage and 2.3 percentage compared to TextRank and LexRank algorithm respectively.

In order to further enhance the summary quality in the context of multidocument summarizing, we would like to investigate more strategies in the future, such as methods based on reinforcement learning. We also want to use our approach for additional tasks, such as answering multidocument questions.

## REFERENCES

- [1] G. Eason, Endres-Niggemeyer, Brigitte, Summarizing Information: Including CD-ROM 'SimSum', Simulation of Summarizing, for Macintosh and Windows. Springer Science & Business Media, 2012.
- [2] Salton, Gerard, et al. 'Automatic analysis, theme generation, and summarization of machine-readable texts', Information retrieval and hypertext. Springer US, 1996. 51-73.
- [3] Bornmann L, Mutz R (2015) Growth rates of modern science: a bibliometric analysis based on the number of publications and cited references. Journal of the Association for Information Science and Technology 66(11):2215.
- [4] Atanassova I, Bertin M, Larivièrè V (2016) On the composition of scientific abstracts. Journal of Documentation.
- [5] Cohan A, Goharian N (2018) Scientific document summarization via citation contextualization and scientific discourse. Int J Digit Libr 19(2-3):287.
- [6] Hahn, Udo, Inderjeet Mani, 'The challenges of automatic summarization', Computer 33.11, 2000, page: 29-36.
- [7] Barzilay, Regina, Michael Elhadad, 'Using lexical chains for text summarization', Advances in automatic text summarization, 1999, page: 111-121.
- [8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in Advances in neural information processing systems, 2017, pp. 5998–6008.
- [9] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," arXiv preprint arXiv:1910.10683, 2019.
- [10] A. See, P. J. Liu, and C. D. Manning, "Get to the point: Summarization with pointer-generator networks," arXiv preprint arXiv:1704.04368, 2017.
- [11] I. Cachola, K. Lo, A. Cohan, and D. S. Weld, "Tldr: Extreme summarization of scientific documents," arXiv preprint arXiv:2004.15011, 2020.
- [12] S. MacAvaney, S. Sotudeh, A. Cohan, N. Goharian, I. Talati, and R. W. Filice, "Ontology-aware clinical abstractive summarization," in Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2019, pp. 1013–1016.
- [13] O. Bodenreider, "The unified medical language system (umls): integrating biomedical terminology," Nucleic acids research, vol. 32, no. suppl 1, pp. D267–D270, 2004.
- [14] C. P. Langlotz, "Radlex: a new method for indexing online educational materials," pp. 1595–1597, 2006.
- [15] S. Sotudeh, N. Goharian, and R. W. Filice, "Attend to medical ontologies: Content selection for clinical abstractive summarization," arXiv preprint arXiv:2005.00163, 2020.
- [16] J. Liang, C.-H. Tsou, and A. Poddar, "A novel system for extractive clinical note summarization using ehr data," in Proceedings of the 2nd Clinical Natural Language Processing Workshop, 2019, pp. 46–54.
- [17] W.-H. Weng, Y.-A. Chung, and S. Tong, "Clinical text summarization with syntax-based negation and semantic concept identification," arXiv preprint arXiv:2003.00353, 2020.
- [18] A. E. Johnson, T. J. Pollard, L. Shen, H. L. Li-Wei, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark, "Mimic-iii, a freely accessible critical care database," Scientific data, vol. 3, no. 1, pp. 1–9, 2016.
- [19] Gong Y, Liu X (2001) Generic text summarization using relevance measure and latent semantic analysis. In: Proceedings of the 24th annual international ACM SIGIR conference on research and development in information retrieval. ACM, pp 19–25.
- [20] Ma T, Nakagawa H (2013) Automatically Determining a Proper Length for Multi-document Summarization: A Bayesian Nonparametric Approach. In: Proceedings of the 2013 conference on empirical methods in natural language processing, pp 736–746.
- [21] Saini N, Saha S, Chakraborty D, Bhattacharyya P (2019) Extractive single document summarization using binary differential evolution: optimization of different sentence quality measures. PLoS One 14(11).
- [22] Louis A, Joshi A, Nenkova A (2010) Discourse Indicators for Content Selection in Summarization. In: Proceedings of the 11th annual meeting of the special interest group on discourse and dialogue. Association for Computational Linguistics, pp 147–156.
- [23] Erkan G, Radev DR (2004) Lexrank: graph-based lexical centrality as salience in text summarization. Journal of Artificial Intelligence Research 22:457.
- [24] Mihalcea R, Tarau P (2004) TextRank: Bringing order into text. In: Proceedings of the 2004 conference on empirical methods in natural language processing, pp 404–411.
- [25] Teufel S, Moens M (2002) Summarizing scientific articles: experiments with relevance and rhetorical status. Computational Linguistics 28(4):409.
- [26] Elkiss A, Shen S, Fader A, Erkan G, States D, Radev D (2008) Blind men and elephants: what do citation summaries tell us about a research article? Journal of the American Society for Information Science and Technology 59(1):51.
- [27] Hernández-Alvarez M, Gomez JM (2016) Survey about citation context analysis: tasks, techniques, and resources. Nat Lang Eng 22(3):327.
- [28] Hoang CDV, Kan MY (2010) Towards automated related work summarization. In: Proceedings of the 23rd international conference on

- computational linguistics: posters. Association for Computational Linguistics, pp 427–435.
- [29] Cohan A, Soldaini L, Goharian N (2015) Matching citation text and cited spans in biomedical literature: a search-oriented approach. In: Proceedings of the 2015 conference of the North American Chapter of the association for computational linguistics: human language technologies, pp 1042–1048.
- [30] Jaidka K, Chandrasekaran MK, Rustagi S, Kan MY (2016) Overview of the CL-SciSumm 2016 shared task. In: Proceedings of the joint workshop on bibliometric-enhanced information retrieval and natural language processing for digital libraries (BIRNDL), pp 93–102.
- [31] Qazvinian V, Radev DR (2008) Scientific paper summarization using citation summary networks. In: Proceedings of the 22nd international conference on computational linguistics, vol 1. Association for Computational Linguistics, pp 689–696.
- [32] Li L, Mao L, Zhang Y, Chi J, Huang T, Cong X, Peng H (2016) Cist system for cl-scisumm 2016 shared task. In: Proceedings of the joint workshop on bibliometric-enhanced information retrieval and natural language processing for digital libraries (BIRNDL), pp 156–167.
- [33] Nomoto T (2016) NEAL: A neurally enhanced approach to linking citation and reference. In: Proceedings of the joint workshop on bibliometric-enhanced information retrieval and natural language processing for digital libraries (BIRNDL), pp 168–174.
- [34] Klampfl S, Rexha A, Kern R (2016) Identifying Referenced Text in Scientific Publications by Summarisation and Classification Techniques. In: Proceedings of the joint workshop on bibliometric-enhanced information retrieval and natural language processing for digital libraries (BIRNDL), pp 122–131.
- [35] Moraes L, Baki S, Verma R, Lee D (2016) Identifying referenced text in scientific publications by summarisation and classification techniques. In: Proceedings of the joint workshop on bibliometric enhanced information retrieval and natural language processing for digital libraries (BIRNDL), pp 113–121.
- [36] Cortes C, Vapnik V (1995) Support-vector networks. *Machine learning* 20(3):273.
- [37] Conroy J, Davis S (2015) Vector space models for scientific document summarization. In: Proceedings of the 1st workshop on vector space modeling for natural language processing, pp 186–191.
- [38] Malenfant B, Lapalme G (2016) RALI system description for CL-SciSumm 2016 shared task. In: Proceedings of the joint workshop on bibliometric-enhanced information retrieval and natural language processing for digital libraries (BIRNDL), pp 146–155.
- [39] Cao Z, Li W, Wu D (2016) Polyu at cl-scisumm 2016. In: Proceedings of the joint workshop on bibliometric-enhanced information retrieval and natural language processing for digital libraries (BIRNDL), pp 132–138.
- [40] Wan X, Yang J, Xiao J (2007) Manifold-Ranking Based Topic-Focused Multi-Document Summarization. In: *IJCAI*, vol 7, pp 2903–2908.
- [41] Li L, Zhang Y, Mao L, Chi J, Chen M, Huang Z (2017) Cist@ clscisumm-17: multiple features based citation linkage classification and summarization.
- [42] Abura'ed A, Chiruzzo L, Saggion H, Accuosto P, Bravo Serrano `A (2017) Lastus/taln@ Clscisumm-17: cross-document sentence matching and scientific text summarization systems.
- [43] Bird S, Dale R, Dorr BJ, Gibson B, Joseph MT, Kan MY, Lee D, Powley B, Radev DR, Tan YF (2008) The acl anthology reference corpus: a reference dataset for bibliographic research in computational linguistics.
- [44] Jaccard P (1901) ETude comparative de la distribution florale dans une portion des alpes et des jura. *Bull Soc Vaudoise Sci Nat* 37:547.
- [45] Blei DM, Ng AY, Jordan MI (2003) Latent dirichlet allocation. *Journal of machine Learning research* 3(Jan):993.
- [46] Lauscher A, Glavas G, Eckert K (2017) University of Mannheim@ CLSciSumm-17: Citation-based summarization of scientific articles using semantic textual similarity. *CEUR workshop proceedings* 2002:33–42. RWTH.
- [47] Dipankar Das S, Pramanick A (2017) In: Proc. of the 2nd joint workshop on bibliometric-enhanced information retrieval and natural language processing for digital libraries (BIRNDL2017). Tokyo, Japan (August 2017).
- [48] Karimi S, Moraes L, Das A, Shakery A, Verma R (2018) Citance based retrieval and summarization using ir and machine learning. *Scientometrics* 116(2):1331.
- [49] Lv Y, Zhai C (2009) Positional language models for information retrieval. In: Proceedings of the 32nd international ACM SIGIR conference on research and development in information retrieval, pp 299–306.
- [50] Tian R, Miyao Y, Matsuzaki T (2014) Logical inference on dependency-based compositional semantics. In: Proceedings of the 52nd annual meeting of the association for computational linguistics (volume 1: Long Papers), pp 79–89.
- [51] Blitzer J, McDonald R, Pereira F (2006) Domain adaptation with structural correspondence learning. In: Proceedings of the 2006 conference on empirical methods in natural language processing. Association for Computational Linguistics, Sydney, pp 120–128. <https://www.aclweb.org/anthology/W06-1615>.
- [52] Cohan A, Goharian N (2017) Scientific article summarization using citation-context and article's discourse structure. arXiv:1704.06619.
- [53] AbuRa'ed A, Bravo Serrano A`, Chiruzzo L, Saggion H (2019) LaSTUS-TALN+ INCO@ CL-SciSumm 2019. BIRNDL@ SIGIR, 224–232.
- [54] Ma S, Xu J, Wang J, Zhang C (2017) NJUST @ CLSciSumm-17. BIRNDL@SIGIR.
- [55] Ma S, Zhang H, Xu J, Zhang C (2018) NJUST@CLSciSumm-18. BIRNDL@SIGIR.
- [56] Debnath D, Achom A, Pakray P (2018) NLP-NITMZ@ CLScisumm-18. BIRNDL@ SIGIR. pp 164–171.
- [57] Zerva C, Nghiem MQ, Nguyen NT, Ananiadou S (2020) Cited text span identification for scientific summarisation using pre-trained encoders. *Scientometrics* 125(3):3109–3137.
- [58] S. Gholamrezazadeh, M. A. Salehi, and B. Gholamzadeh, "A comprehensive survey on text summarization systems," in *Proc. 2nd Int. Conf. Comput. Sci. Appl.*, Dec. 2009, pp. 1–6.
- [59] R. Mishra, J. Bian, M. Fiszman, C. R. Weir, S. Jonnalagadda, J. Mostafa, and G. D. Fiol, "Text summarization in the biomedical domain: A systematic review of recent research," *J. Biomed. Informat.*, vol. 52, Dec. 2014, pp. 457–467.
- [60] C. Saranyamol and L. Sindhu, "A survey on automatic text summarization," *Int. J. Comput. Sci. Inf. Technol.*, vol. 5, no. 6, 2014, pp. 7889–7893.
- [61] N. Andhale and L. A. Bewoor, "An overview of text summarization techniques," in *Proc. Int. Conf. Comput. Commun. Control Autom. (ICCUBEA)*, Aug. 2016, pp. 1–7.
- [62] M. Gambhir and V. Gupta, "Recent automatic text summarization techniques: A survey," *Artif. Intell. Rev.*, vol. 47, no. 1, pp. 1–66, Jan. 2017.
- [63] W. S. El-Kassas, C. R. Salama, A. A. Rafea, and H. K. Mohamed, "Automatic text summarization: A comprehensive survey," *Expert Syst. Appl.*, vol. 165, Mar. 2021, Art. no. 113679.
- [64] Li L, Zhu Y, Xie Y, Huang Z, Liu W, Li X, Liu Y (2019) CIST@ CLSciSumm-19: Automatic Scientific Paper Summarization with Citances and Facets. In: BIRNDL@ SIGIR, pp 196–207.
- [65] La Quatra M, Cagliero L, Baralis E (2019) Poli2Sum@ CLSciSumm-19: Identify, Classify, and Summarize Cited Text Spans by means of Ensembles of Supervised Models. In: BIRNDL@ SIGIR, pp 233–246.
- [66] Ma S, Zhang H, Xu T, Xu J, Hu S, Zhang C (2019) IR&TMNJUST @ CLSciSumm-19. In: BIRNDL@ SIGIR, pp 181–195.
- [67] Chiruzzo L, AbuRa'ed A, Bravo A`, Saggion H (2019) LaSTUSTALN+INCO@ CL-SciSumm 2019. BIRNDL@ SIGIR, pp 224–232.
- [68] R. Senthamizh Selvan, Dr. K. Arutchelvan. (2021). An Effective Approach for Abstractive Text Summarization using Semantic Graph Model. *Annals of the Romanian Society for Cell Biology*, 13925–1393.

# Convolutional Neural Networks with Transfer Learning for Pneumonia Detection

Orlando Iparraguirre-Villanueva<sup>1</sup>, Victor Guevara-Ponce<sup>2</sup>, Ofelia Roque Paredes<sup>3</sup>, Fernando Sierra-Liñan<sup>4</sup>,  
Joselyn Zapata-Paulini<sup>5</sup>, Michael Cabanillas-Carbonell<sup>6</sup>

Facultad de Ingeniería y Arquitectura, Universidad Autónoma del Perú, Lima, Perú<sup>1</sup>

Escuela de Posgrado, Universidad Ricardo Palma, Lima, Perú<sup>2,3</sup>

Facultad de Ingeniería, Universidad Privada del Norte, Lima, Perú<sup>4</sup>

Escuela de Posgrado, Universidad Continental, Lima, Perú<sup>5</sup>

Escuela Técnica Superior de Ingenieros de Telecomunicación, Universidad Politécnica de Madrid<sup>6</sup>

Vicerrectorado de Investigación, Universidad Norbert Wiener, Lima, Perú<sup>6</sup>

**Abstract**—Pneumonia is a type of acute respiratory infection caused by microbes, and viruses that affect the lungs. Pneumonia is the leading cause of infant mortality in the world, accounting for 81% of deaths in children under five years of age. There are approximately 1.2 million cases of pneumonia in children under five years of age and 180 000 died in 2016. Early detection of pneumonia can help reduce mortality rates. Therefore, this paper presents four convolutional neural network (CNN) models to detect pneumonia from chest X-ray images. CNNs were trained to classify X-ray images into two types: normal and pneumonia, using several convolutional layers. The four models used in this work are pre-trained: VGG16, VGG19, ResNet50, and InceptionV3. The measures that were used for the evaluation of the results are Accuracy, recall, and F1-Score. The models were trained and validated with the dataset. The results showed that the InceptionV3 model achieved the best performance with 72.9% accuracy, recall 93.7%, and F1-Score 82%. This indicates that CNN models are suitable for detecting pneumonia with high accuracy.

**Keywords**—Neural networks; transfer learning; pneumonia; detection; Convolutional

## I. INTRODUCTION

Pneumonia is a type of acute respiratory infection caused by bacteria, viruses, or fungi that affects the lungs. Pneumonia is the leading cause of child mortality worldwide, with pneumonia killing an estimated 920136 children under five years of age in 2015, accounting for 15% of all deaths in children under five years of age worldwide [1]. Pneumonia is more prevalent in underdeveloped countries, where the lack of basic conditions is notorious, pollution worsens the situation and, at the same time, medical resources are increasingly scarce [2]. Therefore, early diagnosis and treatment play a very important role in preventing the disease from becoming fatal. Chest radiographs (X-rays) are most often used for the diagnosis of pneumonia. However, X-rays are prone to subjective variability [3][4]. Therefore, this paper presents four CNN models to detect pneumonia from chest X-ray images, which can be used by medical centers to detect pneumonia in their patients. The four deep CNN models were trained to classify the X-ray images into two types: normal and pneumonia.

The CNN-based transfer learning (TL) models used for this research work are: VGG16, VGG19, ResNet50, and InceptionV3, these models have been trained with the ImageNet database, a database with millions of images, and have obtained very satisfactory results, considered as successful. However, for this work, we used the Kaggle dataset, which is a less extensive dataset, basically due to computational resources. The four classification models were developed using CNNs for the purpose of detecting pneumonia from chest X-ray images. It is important to note that none of the four models used in this work have the same number of convolutional layers, so there is no direct relationship with the accuracy of the model [5]. Therefore, each of the models delivers different results with respect to accuracy. Each of the models follows its training architecture. To obtain the best accuracy in each model, at first, it is trained with a certain amount of convolutional combinations, dense layers, dropout, and other optimizers evaluating each model after each iteration, later, the complexity was increased to obtain a better model accuracy. The aim of this work is to classify and detect pneumonia from chest X-ray images, using CNNs with TL. If the model manages to achieve high accuracy, but with low recall values, it is considered an unreliable, ineffective, or even unsafe performance, since high false negative values represent a higher number of cases where the model predicts an output as normal, but in reality, the output is not normal. Therefore, to avoid this, we consider only the models with the highest recall and accuracy values [6] [7]. This is why, in the case of medical image processing, retrieval is preferred over other performance evaluation parameters.

This paper is organized as follows: Section 1 provides an introduction to the subject, addressing the problems, importance, purpose, and objective for undertaking this work. Section 2 explores the main works related so far. Section 3 describes the work methodology, the architectures of the models to be trained, the process diagram, and the data set to be trained. Section 4 presents the results obtained by the four CNN models and discusses the results. Finally, Section 5 provides the conclusions of the work.



## II. RELATED WORK

Academic researchers and scientists in the medical sciences have published research papers addressing the problem of pneumonia detection with neural networks.

The authors in [5] presented a CNN model to detect pneumonia with high accuracy from chest images. The experiment worked with chest X-ray images, achieving an accuracy of 89.67%. Similarly, in [8][9] they developed work to automatically detect bacterial pneumonia, through the TL, for which they used 5247 chest X-ray images. Classifying in three groups: normal, bacterial, and viral pneumonia, obtaining results in classification accuracy of 98%, 95%, and 93%, respectively. Similarly, in [10] they developed four models: CNN, VGG16, VGG19, and InceptionV3, where they used TL techniques with CNN. They used 9992 normal chest radiographs and 2972 pneumonia. The results were tested with 854 pneumonia and 849 normal chest images, obtaining an accuracy higher than 97% in all four models. As well as, in [11] applied an automated TL approach with CNN using four pre-trained models (VGG19, DenseNet121, Xception, and ResNet50), to identify pneumonia. the performance of the four models provided an accuracy higher than 83.0%. Also, in [12] they trained different CNNs to classify X-ray images into two types, normal and pneumonia. As well as, in [13] they proposed a learning framework combining residual thinking and convolution to diagnose childhood pneumonia.

CNNs have the function of automatically extracting features. Currently, it is the area of choice for disease diagnosis related to radiographic image analysis, the purpose of which is to aid in the early detection of symptoms and risk factors of the Covid-19 virus. Also, [14] explored the effectiveness of AI to quickly and accurately identify COVID-19 from a set of images with different deep learning (DL) models and adjust them to achieve better detection accuracy of the best algorithms. Also, in [15] they evaluated the performance of CNN architectures for X-ray image classification, concluding that the procedure called TL produces the best results for the detection of various anomalies. Similarly, in [16] they proposed a DL technique based on object detection, CNN, and TL, combined with the pre-trained VGG19 model. For this, they used 1583 healthy samples and 5273 with pneumonia. The proposed model achieved an accuracy of over 99%.

## III. METHODOLOGY

Convolution is the main element in the construction of a CNN. The convolution is composed of several layers of convolutional filters, to which activation functions and optimizers are added to achieve a better result. The work involves a series of steps, starting with the import of the dataset from Kaggle, and then the processing of the dataset is performed. Next, the dataset was trained with the following models: VGG16, VGG19, ResNet50, and Inception-v3 for classification. A total of 5216 chest X-ray images with Pneumonia were used for training and 624 for validation. The following sections describe the above steps in more detail. The phases of this work are shown in Fig. 1.

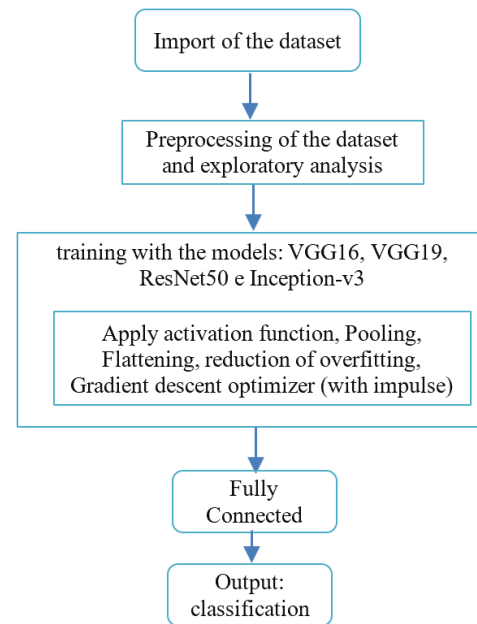


Fig. 1. Diagram of the Process that Follows the Work.

### • Dataset Processing

For preprocessing, a total of 5840 images were used, divided into 5216 chest X-ray images for training and 624 for experimental validation. At this stage, we assigned the parameters and standardized the channel numbers, image dimensions, Batch size, validation size, regulation, and early stopping techniques, and applied the data for training and validation as shown in Fig. 2.

### • Training with the Models

The images are classified into two types: normal and pneumonia, as shown in Fig. 3. In addition, fine-tuning is applied to match the outputs to the classes according to the problem. This consists of four densely connected layers; the first and second are assigned 4096 neurons respectively, the third is reduced to 1000 neurons, and finally, two neurons are used for the output, one for each class. Then the function to generate the model is applied, and with that, the training is started by defining the number of epochs, the batch size, the number of images to train, the number of test images, and the validation steps.

```
# Training
train_image_cogenerator = image_generator.flow_from_directory(
    train_path,
    target_size=image_shape[:2],
    color_mode='rgb',
    batch_size=batch_size,
    class_mode='categorical')

# Test
test_image_cogenerator = image_generator.flow_from_directory(
    test_path,
    target_size=image_shape[:2],
    color_mode='rgb',
    batch_size=batch_size,
    class_mode='categorical')
```

Fig. 2. Training and Test Code.



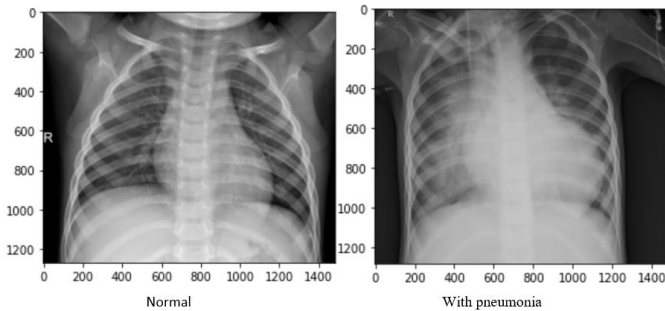


Fig. 3. Chest X-Ray Images Include Two Types of Images: Normal and with Pneumonia.

#### A. CNN Architecture

CNN is a type of multilayer artificial neural network. As shown in Fig. 4, it is composed of four convolutional layers and among other reduction layers, these are assigned alternately, and at the end, total connection layers are added, similar to a multilayer perceptron network. In the following, each of the layers is discussed.

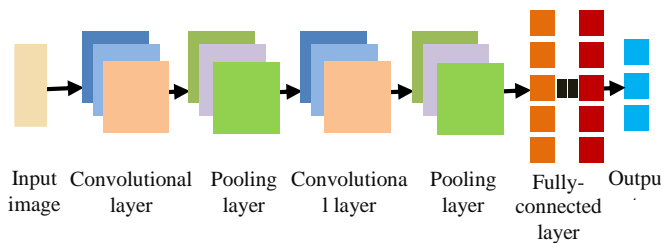


Fig. 4. CNN Architecture.

#### B. Convolutional Layer

The convolutional layer is in charge of processing the output of the neurons that are connected in the local input areas, calculating the product to be scaled between their weights. One of the functions of this layer is to reduce the dimensions of the images, facilitating their processing. However, this leads to the loss of some information, but the integral part of the image features is preserved since it is retained by the detector [17]. Also, in this layer filters, and attribute detectors were applied to the input matrix. Then a clustering process is applied to obtain our first convolutional layer.

#### C. Activation Functions

The activation function (AF) is the one that returns an output that is generated by the neuron given an input. Each of the layers that make up a neural network has an AF that allows prediction. The AF is divided into two types: linear and nonlinear. The linear function allows the input data to be equal to the output data, it is also applied when linear regression is required as output. Meanwhile, the nonlinear function is applied when you want to classify or when you have categorical outputs.

#### D. Pooling Layer

This layer is applied between the two convolution layers, as input, it receives the feature map formed at the output of the convolution layer; its main function is to reduce the size of the

images while preserving their most resolving features [18]. Finally, in the output of each Pooling layer, the same number of features is obtained as in the output, but considerably compressed [16].

The most commonly used clustering techniques with the different models are Max Pooling and Average Pooling. Also, max Pooling is used to create a feature map with reduced sampling. Average Pooling is used to calculate the average value of the filter size. The application of these two pooling techniques provides the ability to learn invariant features and also acts as a regularizer to reduce the overfitting problem. In addition, they significantly reduce the computational cost and training time of the networks, which are important criteria to consider.

#### E. Fully Connected Layer

This layer allows the feature maps generated from the neural network to be processed in a very facial way. Next, the image to be trained (input) passes through the convolution and clustering layer then enters the fully connected (FC) layer. In this way, the input image continues forward by calculating the weights. An FC neural network is composed of a set of FC layers and, connects to each neuron in a layer. The FC layer also functions as a classifier in the CNN. This layer has a behavior similar to a traditional network. For this case study, the FC layer is implemented through a convolutional operation. The FC layer is FC to the previous layer and, the convolution layer is used to replace the FC layer. Usually, a 1x1 convolution kernel is used. This type of CNN does not include an FC layer; thus, it can be converted into a full convolution of a neural network [19]. In Fig. 2, the first two layers are used to manage the features, while the last two layers are FC, and are used for classification.

#### F. Reducing Overfitting

The dropout technique was applied to reduce the overfitting in the VGG16, VGG19, and ResNet50 models. Dropout is a regulation technique based on the elimination of neurons in the neural network layers that are applied based on the probability given by the distribution [20]. The main objective of this technique is to mitigate the possible occurrence of phenomena known as overfitting. This phenomenon is very characteristic of neural networks and occurs in most training. There are multiple ways to reduce overfitting. Increasing the volume of data is one way to reduce overfitting, however, this leads to an increase in computational resources and training time.

#### G. Transfer Learning

With new advances in CNNs, the margins of error in image classification have been reduced. Models such as ResNet and DenseNet were developed to improve image classification, achieving exceptional performance in large-scale visual recognition. In recent years, TL has been used very successfully in different fields, such as manufacturing processes, medicine, and baggage screening, among others [21][8]. This has allowed eliminating the requirement of large amounts of data for training, it also allows reducing the training time, hence less computational resources. Fig. 5 shows the model from large volumes of data, and how it can be used for smaller data.

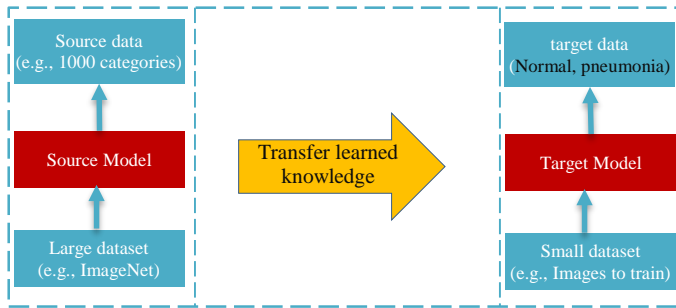


Fig. 5. Transfer Learning Behavior.

H. Model Architecture

Four models were trained with the normal and pneumonia images used in this work. The following is a detailed description of the models presented in this work.

1) *VGG16*: VGG16 is a CNN model that achieved 92.7% accuracy among the top five in the ImageNet dataset containing more than 14 million images divided into 1000 categories [22]. The main reasons for using this model are a) its architecture is easy to understand as well as its implementation; b) it contains relatively few convolutional layers: 13 convolutional and 3 dense layers; c) it has been trained with the ImageNet database.

This model starts with an input image of dimensions (224,224,3), as shown in Fig. 6. The first two layers are composed of 64 channels with a filter size of (3,3), then, after a layer (2,2), there are two convolutional layers with 256 channels and with a filter size of (3,3). After that, there are two sets of three convolution layers and a maximum group layer, with 512 channels and filter size (3,3). And so on, successively, until the dense layers are reached.

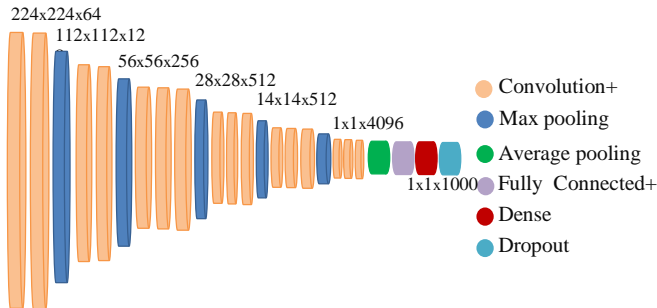


Fig. 6. Architecture of VGG16.

2) *VGG19*: The VGG19 model is very similar to VGG16, they follow the same logic, with the difference that VGG19 has a greater number of layers, but the objective of both models is common: to filter the image keeping only the discriminant information. The VGG19 model has five blocks as shown in Table I. The first two contain two convolutional layers with filter sizes of 64 and 128; the middle block contains four convolutional layers with a filter size of 256, and the last two contain two convolutional layers with a filter size of 512 each [23].

This model, in its first phase, freezes the entire network, except for the fully connected layer. In a very similar way, as in the VGG16 model, the same connected layer has been used in the VGG19 model. Then, in the second layer, the last four convolutional layers are unfrozen to learn new features. The architecture of the model is shown in Table I.

TABLE I. VGG19 MODEL ARCHITECTURE

Type	N° Filters / Parameters			
2 Conv2D	64	64		
Max Pool	N/A			
2 Conv2D	128	128		
Max Pool	N/A			
4 Conv2D	256	256	256	256
Max Pool	N/A			
4 Conv2D	512	512	512	512
Max Pool	N/A			
4 Conv2D	512	512	512	512
Max Pool	N/A			
3 layers Fully-Connect	4096	4096	1000	
Softmax	N			

3) *ResNet50*: ResNet is classified in the residual networks category. It was developed by Microsoft Research and managed to win first place in the IRSVRC 2015 competition, obtaining a 3.57% margin of error in the top five [12]. It is mainly used for image classification. In this type of network, instead of waiting for the layers to conform to the desired mapping, it is left to conform to a residual mapping. Residual learning is adopted for every certain number of stacked layers. Formally, a building block described in equation 1.1 is considered.

$$y = F(x, w_i) + x \tag{1.1}$$

Where x is the input y, y is the output of the considered layer group, and F (X, Wi) represents the residual mapping to be learned. There are two types of blocks for residual networks [24]. The building block is composed of two convolutional layers with a 3x3 size filter; and the bottleneck building block is composed of three convolutional layers, the first and third layers with 1x1 filters, the second with a 3x3 filter as can be seen in Fig. 7.

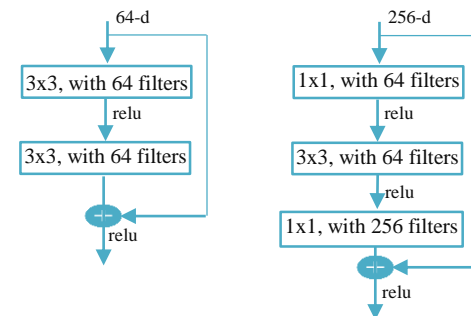


Fig. 7. Representation of the Building Block and the Bottleneck Building Block.

4) *Inception-v3*: Inception -v3, is a deep CNN with 42 layers, developed by Google, has a high image classification performance, and is integrated by: convolution layers, avg Pool, MaxPool, Concat, DropOut, Fully Connected and softmax, as shown in Fig. 8. The network has multiple versions from Inception -v1, Inception -v2 and Inception -v4, each version has been presenting significant improvements for

a better adaptation of the model [25]. This version is much more complex to train; it takes more time, even days. However, this problem is solved with TL [26], since the last layer of the model is preserved for the new classes and, the Inception-v3 model is undone by removing the last layer through the TL technique.

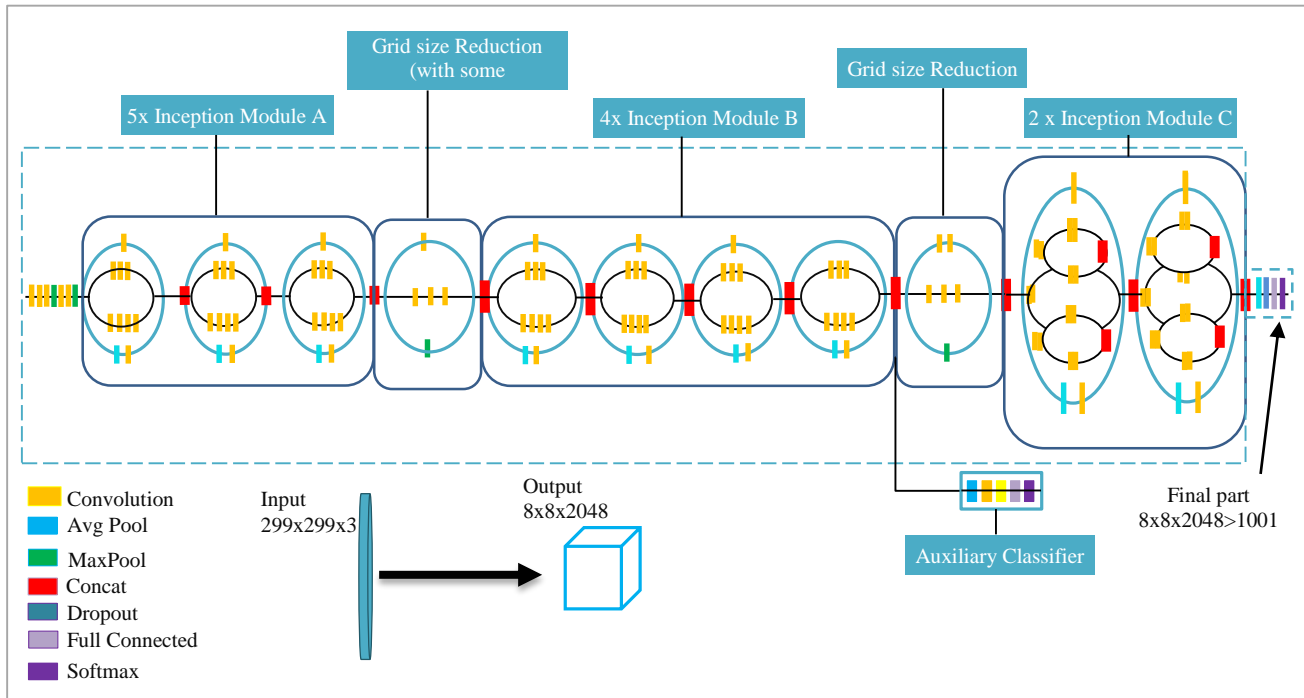


Fig. 8. Architecture of Inception-v3.

#### IV. RESULTS AND DISCUSSIONS

This section presents the results of the four models trained and validated with a dataset of 5840 chest X-ray images with pneumonia and, is organized as follows: 5216 chest X-ray images for training and 624 images for model validation. The same processing technique, the same amount of data, and the same number of partitions were used for all four models. With the predicted values, the confusion matrix was constructed for all models, where the predictions are synthesized and compared with the real values. For example, Fig. 9 shows the confusion matrix of the VGG19 model, where the true positives (correct positive predictions), true negatives (correct negative predictions), false positives (incorrect positive prediction), and false negatives (incorrect negative prediction) can be observed.

Similarly, to measure the performance of each of the models, the Accuracy function was used to evaluate the overall correct predictions (true positives and true negatives) among the total predictions (true positives, false positives, true negatives and false negatives). Loss, this function calculates the incorrectly classified predictions (false positives and false negatives), among the total predictions (true positives, false positives, true negatives, and false negatives). This gives us a clearer picture of how well the model is performing. Recall allows us to determine how many of the predicted positives were found to be correct.

The recall or sensitivity to measure the quality of a machine learning model is extremely important in classification tasks. Recall, which is responsible for calculating the percentage of hits, is also known as sensitivity. F1-Score allows the combining of both Accuracy and Recall into a single weighted measure. If the F1-Score is high, this means that false positives and false negatives are low.

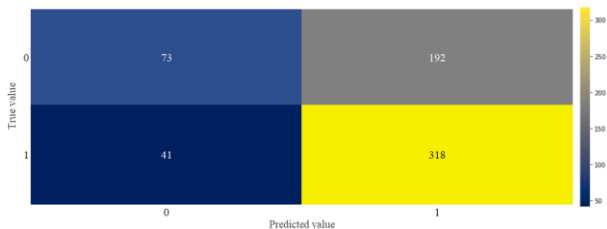


Fig. 9. Confusion Matrix of the VGG19 Model.

Graphs were also constructed to evaluate the performance of each of the models as the epochs developed. As shown in Fig. 10, 11, 12, and 13, the performance of the models presented in this work is analyzed below.

Fig. 10, 11, 12, and 13 show a comparison between Accuracy and loss, i.e. the error with training and validation data. The error with the training data is much smaller as we increase the number of epochs. In contrast, in the validation data, Loss differs a lot from its real value as the number of epochs increases, thus it is very unstable. Regarding the accuracy, the blue line indicates that the accuracy percentage of the model is very close to one as the number of epochs increases, unlike the accuracy in the validation data, where it is very high for some cases.

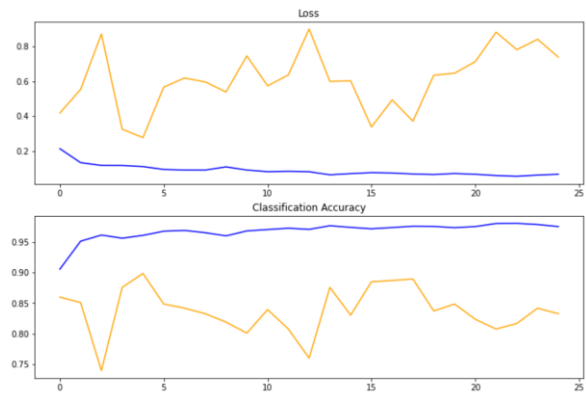


Fig. 10. Accuracy and Data Loss of the VGG16 Model.

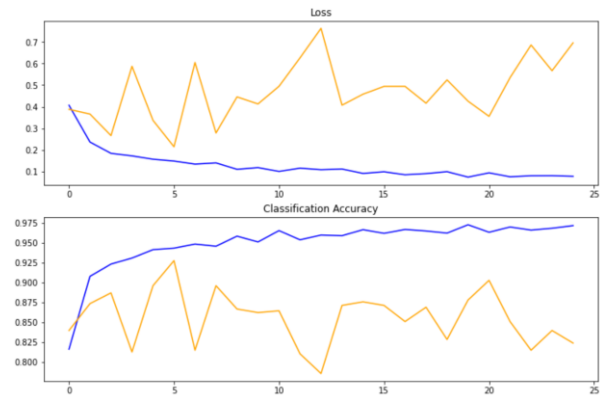


Fig. 11. Accuracy and Data Loss of the VGG19 Model.

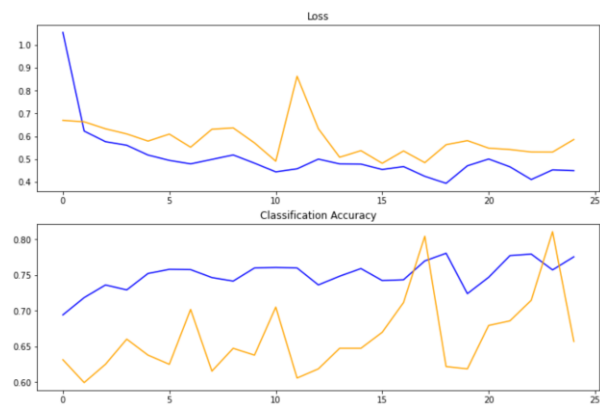


Fig. 12. Accuracy and Data Loss of the ResNet50 Model.

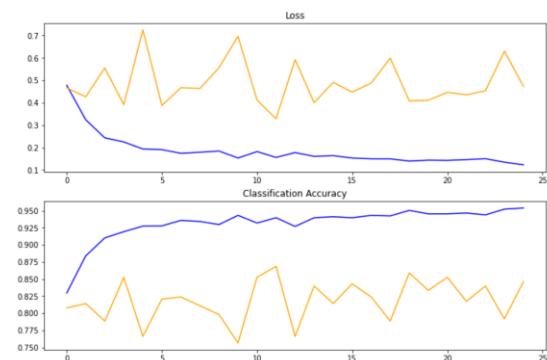


Fig. 13. Accuracy and Loss of Data from Inception-v3.

In the evaluation of the results, the two types of patients were taken into account: normal patients and patients with pneumonia. The confusion matrix in Table II shows the error generated by each model. This helps us to know the performance of each model in the classification of the validation images. Table II shows the confusion matrix of the four CNN models.

TABLE II. CONFUSION MATRIX OF THE MODELS UNDER STUDY

Model	Predicted	
VGG19	71	194
	40	319
VGG19	73	189
	41	321
ResNet50	63	178
	18	365
Inception-v3	86	144
	25	369

Concerning the performance values for each model, the recall and F1-Score are calculated based on the above-described confusion matrices. A comparison of the training and data validation results for each model is presented below.

With the results in Tables III and IV, it is resolved that the VGG16 model in training managed to obtain an accuracy of 62.5% and a loss of 41%, respectively. Similarly, for the VGG19 model, the level of accuracy in training shows a slight improvement, reaching 63.1% and a 41% loss in validation. With the ResNet50 model, the training accuracy shows a significant improvement, reaching 68.6% accuracy, and a 37% loss in validation. Finally, the InceptionV3 model evidences a better performance, achieving 72.9% accuracy in training and a 23% loss in validation. Therefore, it can be concluded that the InceptionV3 model has outperformed the VGG16, VGG19, and ResNet50 models, given that it has achieved the best values in each performance measure, both in accuracy and F1-score, demonstrating that it is a coherent and effective model obtaining a score of over 93% in the recall, although ResNet50 also obtained an excellent recall score of 95.3%. Inceptionv3 is superior in accuracy with 72.9% and F1-Score with 82%. Models VGG16, VGG19, and ResNet50 did not achieve the best score, as can be seen in Fig. 10, 11, and 12, which show the accuracy and loss value of each model.

TABLE III. VALUES OF MODEL PERFORMANCE MEASURES

Model	Accuracy	Recall	F1
VGG16	62.5%	88.9%	73.4%
VGG19	63.1%	88.7%	73.8%
ResNet50	68.6%	95.3%	79.8%
InceptionV3	72.9%	93.7%	82%

TABLE IV. PRECISION AND LOSS VALUES FOR EACH MODEL

Model	Training		Validation	
	Accuracy	loss	Accuracy	Loss
VGG16	70%	31%	62.5%	41%
VGG19	72%	32%	63.1%	39%
ResNet50	69%	36%	68.6%	37%
InceptionV3	83%	24%	72.9%	23%

In the TL models, the confusion matrices represent the margin of error in the classifier models. The results obtained from the training and validation data are shown in Table IV. Note that models VGG16, VGG19, and ResNet50 show relatively high overfitting concerning InceptionV3 because the difference between the accuracy obtained in the training and the accuracy obtained with the validation data is very noticeable. These three models have high losses and their accuracy is also low, therefore, these three models have poor performance, and little efficiency could be improved or trained with more data volume. In general, the VGG19 model performs better than the VGG16 model; the ResNet50 model outperforms the VGG16 and VGG19 models in all metrics (accuracy, recall, and F1-score), respectively. The accuracy obtained in the training and the loss value can be seen in Fig. 10, 11, and 12, where the variations of the accuracies are shown according to the number of epochs that are increased. The models used in this work correspond to deep CNNs, each model with a certain number of layers. Their accuracy can improve with a higher volume of data for training.

## V. CONCLUSION

This paper presents four high-performance CNN models for application in real medical cases. All four models have high training accuracy rates. Recall is an important factor in measuring performance since it is important to reduce the number of false negatives in the image processing to be trained. While the ResNet50 model achieved the best recall of 95.3%, the Inceptionv3 model also achieved a recall of 93.7%, and the two models with the best F1-Score of 79.8% and 82% respectively.

The Inceptionv3 model is the best performing model in terms of accuracy and F1 score; therefore, it is the model that achieved the highest training efficiency. This model can be used by clinicians to detect pneumonia early in both children and adults. This model can be taken to an automated process to process a large number of X-ray images automatically and provide accurate diagnostic results, helping healthcare facilities to provide better patient services and thus reduce the mortality rate from this disease.

For future work, they seek to improve the accuracy of the models with new tuning parameters and optimizers. In [27] they presented a model based on Xception for the detection of breast cancer in real-time. The InceptionV3 model, which is the best performing model in this work, can be extended to classify other diseases as [27] has done with good results. The performance of the models can be improved with a larger amount of data.

#### ACKNOWLEDGMENT

To the PhD program of the Escuela Técnica Superior de Ingenieros de Telecomunicación, Universidad Politécnica de Madrid, Spain.

#### REFERENCES

- [1] Organización Mundial de la Salud, “Neumonía,” 2021. <https://www.who.int/es> (accessed Aug. 24, 2022).
- [2] R. Kundu, R. Das, Z. W. Geem, G. T. Han, and R. Sarkar, “Pneumonia detection in chest X-ray images using an ensemble of deep learning models,” *PLoS One*, vol. 16, no. 9 September, Sep. 2021, doi: 10.1371/JOURNAL.PONE.0256630.
- [3] M. I. Neuman *et al.*, “Variability in the interpretation of chest radiographs for the diagnosis of pneumonia in children,” *J Hosp Med*, vol. 7, no. 4, pp. 294–298, Apr. 2012, doi: 10.1002/JHM.955.
- [4] G. J. Williams *et al.*, “Variability and accuracy in interpretation of consolidation on chest radiography for diagnosing pneumonia in children under 5 years of age,” *Pediatr Pulmonol*, vol. 48, no. 12, pp. 1195–1200, Dec. 2013, doi: 10.1002/PPUL.22806.
- [5] V. Sirish Kaushik, A. Nayyar, G. Kataria, and R. Jain, “Pneumonia Detection Using Convolutional Neural Networks (CNNs),” in *Lecture Notes in Networks and Systems*, vol. 121, Springer, 2020, pp. 471–483. doi: 10.1007/978-981-15-3369-3\_36.
- [6] A. Fourcade and R. H. Khonsari, “Deep learning in medical image analysis: A third eye for doctors,” *J Stomatol Oral Maxillofac Surg*, vol. 120, no. 4, pp. 279–288, Sep. 2019, doi: 10.1016/J.JORMAS.2019.06.002.
- [7] O. Iparraguirre-Villanueva, V. Guevara-Ponce, F. Sierra-Liñan, S. Beltozar-Clemente, and M. Cabanillas-Carbonell, “Sentiment Analysis of Tweets using Unsupervised Learning Techniques and the K-Means Algorithm,” *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 6, pp. 571–578, 2022, doi: 10.14569/IJACSA.2022.0130669.
- [8] T. Rahman *et al.*, “Transfer Learning with Deep Convolutional Neural Network (CNN) for Pneumonia Detection Using Chest X-ray,” *Applied Sciences 2020, Vol. 10, Page 3233*, vol. 10, no. 9, p. 3233, May 2020, doi: 10.3390/AP10093233.
- [9] P. Chhikara, P. Singh, P. Gupta, and T. Bhatia, “Deep convolutional neural network with transfer learning for detecting pneumonia on chest x-rays,” *Advances in Intelligent Systems and Computing*, vol. 1064, pp. 155–168, 2020, doi: 10.1007/978-981-15-0339-9\_13/COVER.
- [10] G. Labhane, R. Pansare, S. Maheshwari, R. Tiwari, and A. Shukla, “Detection of Pediatric Pneumonia from Chest X-Ray Images using CNN and Transfer Learning,” *Proceedings of 3rd International Conference on Emerging Technologies in Computer Engineering: Machine Learning and Internet of Things, ICETCE 2020*, pp. 85–92, Feb. 2020, doi: 10.1109/ICETCE48199.2020.9091755.
- [11] M. Salehi, R. Mohammadi, H. Ghaffari, N. Sadighi, and R. Reiazi, “Automated detection of pneumonia cases using deep transfer learning with paediatric chest X-ray images,” *British Journal of Radiology*, vol. 94, no. 1121, May 2021, doi: 10.1259/BJR.20201263.
- [12] R. Jain, P. Nagrath, G. Kataria, V. Sirish Kaushik, and D. Jude Hemanth, “Pneumonia detection in chest X-ray images using convolutional neural networks and transfer learning,” *Measurement (Lond)*, vol. 165, Dec. 2020, doi: 10.1016/j.measurement.2020.108046.
- [13] G. Liang and L. Zheng, “A transfer learning method with deep residual network for pediatric pneumonia diagnosis,” *Comput Methods Programs Biomed*, vol. 187, p. 104964, Apr. 2020, doi: 10.1016/J.CMPB.2019.06.023.
- [14] M. M. Tareh, N. Zhu, T. A. A. Ali, A. S. Hameed, and M. L. Mutar, “Transfer Learning to Detect COVID-19 Automatically from X-Ray Images Using Convolutional Neural Networks,” *Int J Biomed Imaging*, vol. 2021, 2021, doi: 10.1155/2021/8828404.
- [15] I. D. Apostolopoulos and T. A. Mpesiana, “Covid-19: automatic detection from X-ray images utilizing transfer learning with convolutional neural networks,” *Phys Eng Sci Med*, vol. 43, no. 2, pp. 635–640, Jun. 2020, doi: 10.1007/S13246-020-00865-4/TABLES/6.
- [16] O. Dahmane, M. Khelifi, M. Beladgham, and I. Kadri, “Pneumonia detection based on transfer learning and a combination of VGG19 and a CNN built from scratch,” *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 24, no. 3, pp. 1469–1480, Dec. 2021, doi: 10.11591/IJEECS.V24.I3.PP1469-1480.
- [17] J. Rubin Jonathan, D. Sanghavi, C. Zhao, K. Lee KathyLee, A. Qadir, and M. Xu-Wilson, “Large Scale Automated Reading of Frontal and Lateral Chest X-Rays using Dual Convolutional Neural Networks,” Apr. 2018, doi: 10.48550/arkiv.1804.07839.
- [18] Y. Xu, Z. Jia, Y. Ai, F. Zhang, M. Lai, and E. I. C. Chang, “Deep convolutional activation features for large scale Brain Tumor histopathology image classification and segmentation,” *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, vol. 2015-August, pp. 947–951, Aug. 2015, doi: 10.1109/ICASSP.2015.7178109.
- [19] A. Jain, S. Ratnoo, and D. Kumar, “Convolutional neural network for Covid-19 detection from X-ray images,” *Proceedings - 2021 4th International Conference on Computational Intelligence and Communication Technologies, CCICT 2021*, pp. 100–104, Jul. 2021, doi: 10.1109/CCICT53244.2021.00030.
- [20] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, “Improving neural networks by preventing co-adaptation of feature detectors,” Jul. 2012, doi: 10.48550/arkiv.1207.0580.
- [21] T. Mehta and N. Mehendale, “Classification of X-ray images into COVID-19, pneumonia, and TB using cGAN and fine-tuned deep transfer learning models,” 2021, doi: 10.1007/s42600-021-00174-z.
- [22] J. C. Valero Gómez, A. P. Zúñiga Incalla, and J. C. Clares Perca, “Algoritmos de Deep Learning para la detección de Neumonía en infantes a través de imágenes de radiografías del tórax,” in *Actas del Congreso Internacional de Ingeniería de Sistemas*, 2021, pp. 183–194. doi: 10.26439/ciis2021.5586.
- [23] Z. E. M. C. O. C. L. G. B. G. I. F. V.-L. A. Y. R. Eduardo Díaz Gaxiola, “Estudio comparativo de arquitecturas de CNNs en hojas de Pimiento Morron infectadas con virus”.
- [24] J. C. Valero Gómez, A. P. Zúñiga Incalla, and J. C. Clares Perca, “Algoritmos de Deep Learning para la detección de Neumonía en infantes a través de imágenes de radiografías del tórax,” in *Actas del Congreso Internacional de Ingeniería de Sistemas*, 2021, pp. 183–194. doi: 10.26439/ciis2021.5586.
- [25] J. S. Kumar, S. Anuar, and N. H. Hassan, “Transfer Learning based Performance Comparison of the Pre-Trained Deep Neural Networks,” *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 1, pp. 797–805, 2022, doi: 10.14569/IJACSA.2022.0130193.
- [26] M. Mujahid *et al.*, “Pneumonia Classification from X-ray Images with Inception-V3 and Convolutional Neural Network,” 2022, doi: 10.3390/diagnostics12051280.
- [27] B. S. Abunasser, M. R. J. AL-Hiealy, I. S. Zaqout, and S. S. Abu-Naser, “Breast Cancer Detection and Classification using Deep Learning Xception Algorithm,” *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 7, 2022, doi: 10.14569/IJACSA.2022.0130729.



# A Monadic Co-simulation Model for Cyber-physical Production Systems

Daniel-Cristian Crăciunean  
Computer Science and Electrical Engineering Department  
Lucian Blaga University of Sibiu  
Sibiu, Romania

**Abstract**—The production flexibility required by the new industrial revolutions is largely based on heterogeneous Cyber-Physical Production Systems models that cooperate with each other to perform complex tasks. To accomplish tasks at an acceptable pace, CPPSs should be based on appropriate cooperation mechanisms. To this end a CPPS must be able to provide services in the form of functionalities to other CPPSs, and also to use functionalities of other CPPSs. The cooperation of two CPPS systems is done by co-simulating the two models that allow the partial or total access of the functionalities of one system, by the other system. Requests from one CPPS to another CPPS create connection moments of the two models that can only be performed in certain states of the two models. Also, the answers to these requests create connections between the two models in other subsequent states. Optimal aggregation of the behaviors of the two models, by co-simulation, is essential because otherwise it can lead to very long waiting times and can cause major problems if not done correctly. We will see in this paper that the behavior of such a simulation model can be represented by a category, and the co-simulation of two models can be defined by a monad determined by two adjoint functors between the simulation categories of the two models.

**Keywords**—Models; metamodels; co-simulation; adjoint functors; monads; cyber-physical production systems

## I. INTRODUCTION

The new industrial revolutions ("Industry 4.0", "Industry 5.0"), respond to the needs of individualization of production by the widespread introduction of Cyber-Physical Production Systems (CPPS) as central elements in making production flexible at all levels. In essence, CPPSs are complex systems made up of heterogeneous entities and subsystems that cooperate with each other depending on the context in which they evolve at all levels of production.

CPPSs are composable systems, that is, they are systems of systems. The composition operation is determined by the interactions between the subsystems in all the phases of the life cycle of the production process. CPPSs are of overwhelming importance in the production process because they implement the interaction between physical and cybernetic components in distributed networks and therefore represent the fusion between the real and the virtual world as a whole.

In this context, it is obvious the importance of approaches for the optimal design and implementation of CPPS. Modeling and simulation are the most common approaches in the process of designing these systems. Modeling makes it possible to

simulate and analyze production processes as well as make decisions before the actual construction of the manufacturing line [7] [8]. Also, after the construction of the production line, the models can be used for its optimization and diagnosis.

The primary artifact in the process of designing and implementing a CPPS is the model. In such a model the emphasis falls, most of the time, on the interaction and cooperation between the heterogeneous components of the system, and not on the internal functionality of these components [20]. Therefore, classical approaches, from systems theory, cannot satisfactorily respond to these interaction modeling requirements [9,22]. Our approach is motivated by the finding of a deficit of sufficiently strong mathematical mechanisms to be an adequate support for such models [14]. This vacuum of remarkable results is even more accentuated in the field of coupling simulators in dynamic structure scenarios at the level of states [21].

We will see further in this paper that category theory, which, unfortunately, is not used enough in modeling, provides all the ingredients needed to specify such models. Thus, in section 2 we will specify a categorical model of a CPPS, in section 3 we will specify the behavior of a model through a category we call Model Simulation Category, and in section 4 we will see that the co-simulation of two models can be defined by a monad determined by two adjoint functors between the simulation categories of the two models. The paper ends with conclusions.

## II. CATEGORICAL MODELING OF CPPS

An essential phase of the process of developing a model is the conceptualization of the domain [4]. A model is an artifact recognized by an observer as an abstract representation of a real system [1]. This artifact is a syntactic and semantic specification of the real system from the point of view pursued.

The conceptualization of the modeling domain begins with the identification of the generic atomic concepts of the domain that will represent classes of entities in the modeling domain [5]. The state of these atomic concepts is generally specified by the values of some associated attributes. This process continues with the specification of the interaction rules of these atomic concepts in models. We will consider in the following that the models are structured as graphs that have as nodes atomic concepts and as arcs interactions between these atomic concepts. The elementary components of the modeling domain will become atomic concepts in models. The behavior of

atomic concepts is dependent on the state in which they are and the context in which they evolve, i.e., the graph structure in which they are integrated.

The specification of a model involves two specification mechanisms, namely a model specification paradigm through the hierarchical assembly of components and a mathematical model that mimics the behavior of the system. The first mechanism must reflect the static dimension of the model and the second the behavioral dimension of it. Therefore, both the syntax and the semantics of a model are characterized by two dimensions, namely a static dimension and a behavioral dimension.

### A. The Static Dimension of CPPS Models

To specify the syntax of the static dimension of CPPS models we use the categorical sketch (Fig. 1.) which is defined as a tuple  $\mathcal{S}=(\mathcal{G}, \mathcal{C}(\mathcal{S}))$ , where  $\mathcal{G}$  is a graph with typed nodes and arcs, and  $\mathcal{C}(\mathcal{S})$  is a set of constraints on the elements of the graph that specify in categorical terms the conditions that a model must meet. Thus, the categorical sketch becomes a metamodel of the static dimension of our model.

The atomic concepts identified in the conceptualization phase of the domain become nodes of the graph  $\mathcal{G}$  of this sketch. Thus, the metamodel that specifies the static dimension of CPPS models is the graph  $\mathcal{G}$ , which respects the constraints  $\mathcal{C}(\mathcal{S})$ , and which has as nodes the atomic concepts and as arcs the rules of aggregation of these concepts in models.

Therefore, in the metamodel  $\mathcal{S}$  of the static dimension of CPPS models,  $\mathcal{G}$  is a graph whose nodes specify the types of atomic concepts they represent and whose arcs specifies the types of connections between these concepts.

The  $(\mathcal{S})$  component is defined at the metamodel level by a diagram predicate signature [13,15], which maps a set of predicates to a set of special graphs called shape graph arity. These special graphs, called shape graph arity, are then mapped by a set of functors, called diagrams, to the components of the graph of the sketch and therefore receive the types of these components. The set of models of the sketch  $\mathcal{S}$  is represented by the set of functors defined on the graph  $\mathcal{G}$  of the sketch  $\mathcal{S}$ , with values in the category of sets and functions (Set), and which respects the constraints defined by  $\mathcal{C}(\mathcal{S})$  (Fig. 1.). These models are, from a syntactic point of view, also graphs that inherit from the graph of the sketch the type of components.

The components, of the graph  $\mathcal{G}$  of the sketch  $\mathcal{S}$ , are endowed with attributes that are mapped to data domains. The semantics of the static dimension of a model is given by the graph structure of the model and by the values of the attributes of the atomic components.

In our approach, the set of static models represents the set of states in the behavioral model. Such a state is characterized by the graph structure of the model and the values of the attributes.

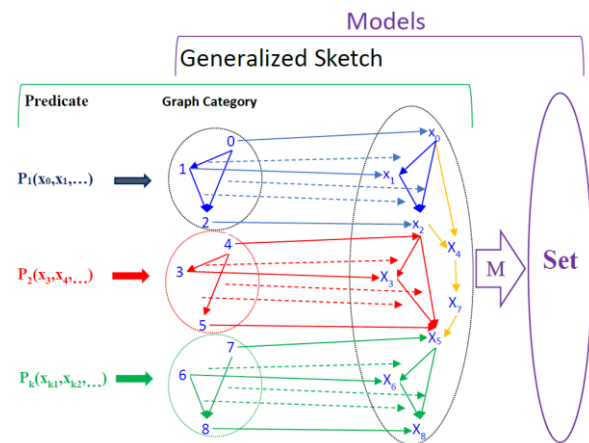


Fig. 1. Generalized Sketch and Models.

### B. The Behavioral Dimension of CPPS Models

The behavior of a CPPS is shaped by the multitude of states in which the CPPS model can be found and by the multitude of transitions that ensure the transition from one state to another. In our approach, the set of states is represented by the set of static models and the set of transitions by a set of transformations of the model that we have called behavioral rules. We associated these behavioral rules with the dynamic components of the model, such as workstation or transport machine components.

We defined the behavioral rules by a tuple  $\tau=(p,\sigma)$ , formed by a graph transformation  $p$  and a behavioral action  $\sigma$ . A behavioral action  $\sigma$ , is a mathematical function that modifies the values of the attributes associated with the graph components in a certain area of the graph structure of the static model. The role of the graph transformation  $p$  is to locate the definition area of the behavioral action and to transform the graph structure from this area of the graph model into another graph structure, provided that the resulting graph structure is also a static model of the categorical sketch.

The syntax of a behavioral rule  $\tau$ , is composed of the syntax of the two components, namely the syntax of the graph transformation  $p$ , and the syntax of the behavioral action  $\sigma$ .

Syntactically a graph transformation is a tuple  $p = (L, R)$ , where  $L$  is a source graph to be transformed and  $R$  is the target graph in which the graph  $L$  is to be transformed. To define the semantics of a graph transformation we used the categorical mechanism called double-pushout (DPO) [16,17,18,19]. The DPO mechanism defines a graph transformation as a span  $p=(L\leftarrow K\rightarrow R)$ , where  $K$  is a common part of the graphs  $L$  and  $R$ , and therefore there are two total inclusion morphisms  $p_L:K\rightarrow L$  and  $p_R:K\rightarrow R$ .

We defined the syntax of a behavioral action by a signature of a behavioral rule that maps a mathematical function to the components of the  $L$  and  $R$  graphs of the graph transformation. Thus, a signature of a behavioral rule is a tuple  $\sigma=(p,C_L,Act,C_R,\alpha)$ , where  $p$  is a graph transformation,  $Act$  is a mathematical function,  $C_L$  is a precondition,  $C_R$  is a postcondition and  $\alpha:\{C_L,Act,C_R\}\rightarrow p$  is an application that

maps the parameters of the Act action and of the preconditions  $C_L$  and  $C_R$  to the nodes of the graphs  $L$  and  $R$ .

The  $L$  and  $R$  components of the graph transformation play in this case the role of shape graphs for the signature of the behavioral action and will be mapped, by means of diagrams, to the graph components of the categorical sketch, mechanism by which they will receive the types of these components. The semantics of behavioral actions will be defined by mathematical functions that recalculate the values of the attributes associated with the graph components  $L$  and  $R$ .

The application of a behavioral rule is preceded by the finding of a total morphism, called a match, from shape graph  $L$  to the static model. If a match is found, the precondition  $C_L$  is checked. If this precondition is verified, the corresponding graph transformation is performed using the DPO algorithm and then the Act action is executed. Finally, the postcondition is checked. If the post-condition is not verified, the behavioral rule cannot be applied and, therefore, we will have to cancel all the effects produced by the partial execution of the behavioral rule. As we can see, a behavioral rule must be applied by an indivisible instruction. We must also note that the behavioral rule must be endogenous, i.e., it must transform a static model of the categorical sketch into another static model of the same sketch. It is obvious that the application of behavioral rules can be done in parallel as long as this application is not conflicting [Ehrig2015].

### III. MODEL SIMULATION CATEGORY

The problem of CPPS optimization is a complex, multi-objective problem, which involves dynamic behavior accompanied by elements of their uncertainty, and therefore cannot be solved by optimization models from classical mathematics. The saving solution to this problem is simulation, which allows the study of the behavior of these complex systems in order to optimize them and eliminate deficiencies from the design phase.

We will understand by simulation the process of imitating the behavior of a materialized system through a multitude of possible trajectories through which it can evolve. Therefore, the simulation can be described as a language on the set of states through which the system can pass. We define an execution of the behavioral model as a word of this language.

In our approach, a state of the behavioral model is represented by a static model of the categorical sketch  $\mathcal{S}=(\mathcal{G},\mathcal{C}(\mathcal{S}))$ , and the transition between these states is made by the behavioral rules. A static model of the sketch  $\mathcal{S}$  is the image of a functor  $\mathfrak{F}:\mathcal{S}\rightarrow\text{Set}$  that maps the nodes of the typed graph  $\mathcal{G}$  of the sketch to sets of components in the category  $\text{Set}$  and the arcs of the graph  $\mathcal{G}$  to functions in  $\text{Set}$ . Next, we will call the image of the sketch  $\mathcal{S}$  by a functor  $\mathfrak{F}$ , instance and we will also denote it with  $\mathfrak{F}$ . The component  $\mathcal{C}(\mathcal{S})$  of the sketch imposes restrictions on the image  $\mathfrak{F}(\mathcal{S})$  in  $\text{set}$ , which will also be respected. Each node  $x$  of the graph  $\mathcal{G}$  represents a type of component of the system, and  $\mathfrak{F}(x)$  is a set of components of type  $x$ . We will consider that these components also contain values for the associated attributes. Also, each arc of the graph  $\mathcal{G}$ , represents an operator of the sketch and is mapped to a

function in the  $\text{Set}$  category. These functions are constitutive elements of the constraints  $(\mathcal{S})$ .

If we have an instance  $\mathfrak{F}_1$  and a behavioral rule  $\tau$ , which transforms  $\mathfrak{F}_1$  into  $\mathfrak{F}_2$ , then  $\tau$  must be endogenous, i.e.,  $\mathfrak{F}_2$ , must also be a static model of the same sketch  $\mathcal{S}$ , this is a condition that must be respected by any behavioral rule. We will denote this by  $\mathfrak{F}_1 \xrightarrow{\tau} \mathfrak{F}_2$ .

With these notations, an execution of a behavioral model, in an initial state  $\mathfrak{F}_0$ , is a chain of behavioral rules:  $\mathfrak{F}_0 \xrightarrow{\tau^0} \mathfrak{F}_1 \xrightarrow{\tau^1} \dots \mathfrak{F}_n \xrightarrow{\tau^n} \dots$ , where  $\mathfrak{F}_k$ ,  $k \geq 0$  are instances and  $\tau^k$ ,  $k \geq 0$  are behavioral rules.

We can introduce a partial operation of composing two behavioral rules. If we have two behavioral rules  $\tau_1=(p_1,\sigma_1)$  and  $\tau_2=(p_2,\sigma_2)$  then the behavioral rule  $\tau=\tau_1 \circ \tau_2$  is defined by composing the components  $(p_1 \circ p_2, \sigma_1 \circ \sigma_2)$ . Since,  $\sigma_1$  and  $\sigma_2$  are mathematical functions, their composition is specific to these functions. For graph transformations, we have a theoretical result that demonstrates that any two sequential graph transformations can be composed into a single equivalent graph transformation that accumulates the effect of both transformations.

Therefore, the set of behavioral rules is endowed with a composition operation. Obviously, this composition operation is associative and to the set of behavioral rules we can add the identity transformation which does nothing. It follows from these considerations that the set of instances together with the set of behavioral rules form a category that we call category of instances and behavioral rules (CIBR). The objects of this category are instances and its arrows are behavioral rules.

Now we can define an execution of a behavioral model, also as the image of the category  $\Omega: 0 \xrightarrow{\alpha_0} 1 \xrightarrow{\alpha_1} \dots k \xrightarrow{\alpha_k} \dots$  through a functor  $\Gamma:\Omega \rightarrow \text{CIBR}$  that specifies the evolution of the model over time:  $\Phi(i)=\mathfrak{F}_i$  for all  $i \geq 0$ ;  $\Phi(\alpha_i)=\tau^i$  for all  $i \geq 0$  as can be seen in Fig. 2. Any execution of a model starting from an initial state  $\mathfrak{F}_0$  produces a simulation trace that is a word in a language defined on the set  $\text{ob}(\text{CIBR})$  of objects of the CIBR category, through the set of behavioral rules.

Thus, if  $\mathfrak{F}=\text{ob}(\text{CIBR})$  then the set of simulation traces, relative to the initial state  $\mathfrak{F}_0$ , form a language  $L(\mathfrak{F}) \subseteq \mathfrak{F}^*$  defined as follows:  $L(\mathfrak{F})=\{ \mathfrak{F}_0 \mathfrak{F}_1 \dots \mathfrak{F}_n \in \mathfrak{F}^* \mid \mathfrak{F}_k = \tau(\mathfrak{F}_j) \text{ for } k \geq 1, k \geq j \geq 0 \text{ and } \tau \text{ is a behavioral rule} \}$ .

If  $L(\mathfrak{F})$  is the language of simulation traces over the vocabulary  $\mathfrak{F}=\text{ob}(\text{CIBR})$ , and  $\beta=\mathfrak{F}_0 \mathfrak{F}_1 \dots \mathfrak{F}_n \in L(\mathfrak{F})$ , we will denote by  $v(\beta)$  the set of instances involved in this trace,  $v(\beta)=\{ \mathfrak{F}_0, \mathfrak{F}_1, \dots, \mathfrak{F}_n \}$ .

We now define a subcategory of the CIBR category, which we call model simulation category (MSC), which has the objects  $\text{ob}(\text{MSC})=\{ \mathfrak{X} \mid (\exists) \beta \in L(\mathfrak{F}) \text{ so that } \mathfrak{X} \in v(\beta) \}$ , and as arrows the corresponding behavioral rules. The PSC category contains all the trajectories on which the CPPS model can evolve. We denote this category by  $\text{MSC}(\mathfrak{F})$ .

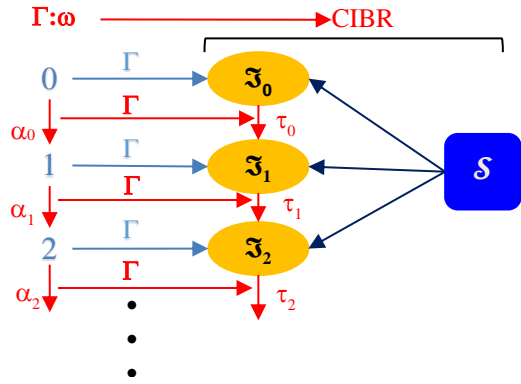


Fig. 2. Execution of a Behavioral Model.

#### IV. MONADIC CO-SIMULATION

The essential role of CPPS in the production processes is to make these processes more flexible in order to realize a lot of individualized products or in small batches, adapted to the customers' requirements. These CPPS must also cover many aspects of the production process. Therefore, there cannot be a single CPPS, which can deal with this variety of requirements. In order to carry out the tasks at an acceptable pace, the CPPS must contain adequate cooperation mechanisms.

A CPPS is a fundamental component of a smart factory and therefore must monitor the status of all participants in the production process and be able to automatically react to any event in order to achieve the objectives. For this purpose, a CPPS must be able to offer services, in the form of functionalities to other CPPSs, and also to use functionalities of other CPPSs.

The cooperation of two CPPS systems is done by co-simulating the two models, which allows partial or total access to the functionalities of one system, by the other system. The co-simulation control is realized by an orchestrator that coordinates the system components according to a co-simulation scenario.

However, in order to achieve a good cooperation, each CPPS must know about the functionalities of the other CPPSs with which it collaborates and how to extract and use the knowledge from the specifications of these CPPSs with which it cooperates in order to achieve the goal. This means that it must encapsulate knowledge about the state and evolution of the other CPPS with which it could cooperate.

The requests of a CPPS to another CPPS create moments of connection of the two models that can only be carried out in certain states of the two models. Also, the answers to these requests create connections between the two models between other subsequent states. This type of interaction between two models is realized between the simulation categories of the two models through co-simulation. The optimal aggregation of the behaviors of the two models, through co-simulation, is essential because, otherwise, it can lead to very long waiting times and can cause major problems if it is not done correctly [25]. We will further introduce a categorical modeling model of this type of connection based on the notion of monad specified by the adjunction of two functors.

On the set  $\mathfrak{S} = \text{ob}(\text{MSC})$ , we introduce a relation  $\preceq$ , defined as follows,  $\mathfrak{S}_i \preceq \mathfrak{S}_j$ , if and only if in the category  $\text{MSC}(\mathfrak{S})$ , there is a path from  $\mathfrak{S}_i$  to  $\mathfrak{S}_j$ , i.e. if  $\mathfrak{S}_j$  can be obtained from  $\mathfrak{S}_i$  through the successive application of a series of behavioral rules. Obviously, this relation is a partial preorder relation and therefore the set  $\mathfrak{S}$  is a preordered set in reference to this relation. We observe that if  $L(\mathfrak{S})$  is the language of simulation traces over the vocabulary  $\mathfrak{S} = \text{ob}(\text{MSC})$ , then all simulation traces  $\beta = \mathfrak{S}_0 \mathfrak{S}_1 \dots \mathfrak{S}_n \in L(\mathfrak{S})$ , respect the condition:  $\mathfrak{S}_i \preceq \mathfrak{S}_j$  if and only if  $i \leq j$ .

We now consider two model simulation categories  $\text{MSC}_1, \text{MSC}_2$  with  $\mathfrak{S} = \text{ob}(\text{MSC}_1), \mathfrak{L} = \text{ob}(\text{MSC}_2)$ . The connection points between the states of the two models, relative to the first model, can be specified by two monotone functions; a request function  $\varphi: \mathfrak{S} \rightarrow \mathfrak{L}$  and a response function  $\psi: \mathfrak{L} \rightarrow \mathfrak{S}$ . To ensure a correct collaboration between the two models, without deadlock situations in the co-simulation flow, it is necessary that any two simulation traces  $\beta_1 = \mathfrak{S}_0 \mathfrak{S}_1 \dots \mathfrak{S}_n \in L(\mathfrak{S})$  and  $\beta_2 = \mathfrak{L}_0 \mathfrak{L}_1 \dots \mathfrak{L}_m \in L(\mathfrak{L})$ , where  $m, n \geq 0$ , to respect the conditions: for each pair of states  $(\mathfrak{S}_i, \mathfrak{L}_j), \mathfrak{S}_i \in v(\beta_1), \mathfrak{L}_j \in v(\beta_2)$  with the property,  $\mathfrak{L}_j = \varphi(\mathfrak{S}_i)$ , there is a state  $\mathfrak{L}_k \in v(\beta_2)$ , such that  $\varphi(\mathfrak{S}_i) \preceq \mathfrak{L}_k$ , and  $\mathfrak{S}_i \preceq \psi(\mathfrak{L}_k)$ . This condition is necessary to avoid deadlock situations, in which the two simulations block each other, because each is waiting for a response from the other.

This type of cooperation between two models in the simulation process requires, therefore, that the functions  $\varphi$  and  $\psi$  form a Galois connection [24]. A Galois connection between the sets  $\mathfrak{S}$  and  $\mathfrak{L}$  is a pair of monotone mappings  $\varphi: \mathfrak{S} \rightarrow \mathfrak{L}$  and  $\psi: \mathfrak{L} \rightarrow \mathfrak{S}$  with the property:  $\varphi(\mathfrak{S}_i) \preceq \mathfrak{L}_j$  if and only if  $\mathfrak{S}_i \preceq \psi(\mathfrak{L}_j)$  where  $\mathfrak{S}_i \in \mathfrak{S}$  and  $\mathfrak{L}_j \in \mathfrak{L}$ . The two applications,  $\varphi$  and  $\psi$  are the left adjunct and, respectively, the right adjunct of the Galois connection.

A theoretical result [23,24] tells us that the condition in the above definition of the Galois connection can be replaced by the condition:  $\mathfrak{S}_i \preceq \psi(\varphi(\mathfrak{S}_i))$  and  $\varphi(\psi(\mathfrak{L}_j)) \preceq \mathfrak{L}_j$ . We will see, next, that this condition has a generalization in category theory and is called adjunction.

Since  $\text{MSC}_1$  and  $\text{MSC}_2$  are categories, we can replace the applications  $\varphi$  and  $\psi$  with two adjoint functors,  $\Phi: \text{MSC}_1 \rightarrow \text{MSC}_2$  and  $\Psi: \text{MSC}_2 \rightarrow \text{MSC}_1$ , and thus obtain an adjunction between two functors, where  $\Phi$  and  $\Psi$  are the left adjunct and, respectively, the right adjunct of the adjunction. The adjunction of two functors  $\Phi$  and  $\Psi$  is denoted by  $\Phi \dashv \Psi$ , where  $\Phi$  is the left adjunct and  $\Psi$  is the right adjunct. The necessary and sufficient condition for two functors to be adjoint is that between the two-variable functors  $\text{Hom}(\Phi, -): \text{MSC}_1 \rightarrow \text{Set}$  and  $\text{Hom}(-, \Psi): \text{MSC}_2 \rightarrow \text{Set}$ , there must be a bijective natural transformation, i.e. we have the natural isomorphism  $\text{Hom}(\Phi, -) \cong \text{Hom}(-, \Psi)$ , where we have denoted with "-" the place of a variable. This condition generalizes the condition from the Galois connection. In other words, in the  $\text{MSC}_1$  and  $\text{MSC}_2$  categories, for each pair of objects  $\mathfrak{S}_p \in \mathfrak{S}$  and  $\mathfrak{L}_q \in \mathfrak{L}$ , there is a behavioral transformation from  $\Phi(\mathfrak{S}_p)$  to  $\mathfrak{L}_q$  in the  $\text{MSC}_2$  category if and only if there is a transformation from  $\mathfrak{S}_p$  to  $\Psi(\mathfrak{L}_q)$  in the  $\text{MSC}_1$  category. But this condition is exactly the necessary and sufficient condition to be able to carry out a

co-simulation without deadlock on all the simulation traces defined by the  $MSC_1$  and  $MSC_2$  categories.

Therefore, the necessary and sufficient condition for the functors  $\Phi$  and  $\Psi$  to form an adjunction is for the natural isomorphism  $\text{Hom}(\Phi(-),-) \cong \text{Hom}(-,\Psi(-))$  to exist. From here it follows that in order to define the functors  $\Phi$  and  $\Psi$  so that they form an adjunction  $\Phi \dashv \Psi$ , it is enough to define a one-to-one relationship between the traces from the categories of simulation models  $MSC_1$  and  $MSC_2$ . This correspondence between the simulation traces is reduced to a bijective natural transformation defined on components  $f: \text{Hom}(\Phi \mathfrak{S}_p, \mathfrak{L}_q) \rightarrow \text{Hom}(\mathfrak{S}_p, \Psi \mathfrak{L}_q)$ , between the behavioral rules of the two behavioral models. Thus if  $\tau_2 \in \text{Hom}(\Phi \mathfrak{S}_p, \mathfrak{L}_q)$  then  $\Psi(\tau_2) = f^{-1}(\tau_2)$ , and if  $\tau_1 \in \text{Hom}(\mathfrak{S}_p, \Psi \mathfrak{L}_q)$ , then  $\Phi(\tau_1) = f(\tau_1)$ . Also, if in the state  $\mathfrak{S}_p \in \mathfrak{S}$  of the  $MSC_1$  model, we have a request to the  $MSC_2$  model, which is in the state  $\mathfrak{L}_r$ , and we need an answer in the  $\mathfrak{S}_r \in \mathfrak{S}$  state, from the  $MSC_2$  model, in the  $\mathfrak{L}_q \in \mathfrak{L}$  state, then we define  $\Phi \mathfrak{S}_p = \mathfrak{L}_r$  and  $\Psi \mathfrak{L}_q = \mathfrak{S}_r$ .

In the context of two adjoint functors we can define two special natural transformations called unit and counit. [10, 12]. For the two adjoint functors  $\Phi \dashv \Psi$ , there is a natural transformation  $\eta: id_1 \rightarrow \Phi \circ \Psi$ , where  $id_1$  is the identity functor from the  $MSC_1$  category, so that for any object  $\mathfrak{S}_k \in \mathfrak{S}$  and  $\mathfrak{L}_1 \in \mathfrak{L}$  and any arrow  $\tau_1^{kt}: \mathfrak{S}_k \rightarrow \Psi(\mathfrak{L}_1) = \mathfrak{S}_t$ , there is a unique arrow  $\tau_2^{tl}: \Phi(\mathfrak{S}_k) = \mathfrak{L}_t \rightarrow \mathfrak{L}_1$  so that the diagram in Fig. 3 commutes [10, 12]. The natural transformation  $\eta$  is called unit adjunction.

Also, the adjunction property of the functors  $\Phi$  and  $\Psi$  assumes the existence of a dual natural transformation  $\varepsilon: \Psi \circ \Phi \rightarrow id_2$ , where  $id_2$  is the identity functor from the category  $MSC_2$ , so that for any arrow  $\tau_2^{tl}: \Phi(\mathfrak{S}_k) = \mathfrak{L}_t \rightarrow \mathfrak{L}_1$ , there is a unique arrow  $\tau_1^{kt}: \mathfrak{S}_k \rightarrow \Psi(\mathfrak{L}_1) = \mathfrak{S}_t$ , so that the diagram in Fig. 4 commutes. The natural transformation  $\varepsilon$  is called counit adjunction.

The adjunction of two functors  $\Phi \dashv \Psi$  is unambiguously specified by the tuple  $(\Phi, \Psi, \eta, \varepsilon)$ , where  $\eta$  is the adjunction unit and  $\varepsilon$  is the adjunction counit.

The adjunction unit  $\eta: id_1 \rightarrow \Phi \circ \Psi$  is also called the insertion of generators and has the role of transforming each object  $\mathfrak{S}_k \in \mathfrak{S}$  into the format of an object  $\Psi(\Phi(\mathfrak{S}_k))$ . This function is executed in the  $MSC_1$  model and can be calculated on components starting from the adjunction condition which says that there is a natural isomorphism  $f: \text{Hom}(\Phi \mathfrak{S}_p, \mathfrak{L}_q) \rightarrow \text{Hom}(\mathfrak{S}_p, \Psi \mathfrak{L}_q)$ . If in this natural isomorphism we make  $\mathfrak{L}_q = \Phi \mathfrak{S}_p$ , then we have the natural isomorphism  $f: \text{Hom}(\Phi \mathfrak{S}_p, \Phi \mathfrak{S}_p) \rightarrow \text{Hom}(\mathfrak{S}_p, \Psi \Phi \mathfrak{S}_p)$  and therefore we can calculate  $\eta$  on the  $\mathfrak{S}_p$  component, according to the formula:  $\eta_{\mathfrak{S}_p} = f(id_{\Phi \mathfrak{S}_p})$  where  $id_{\Phi \mathfrak{S}_p}: \Phi(\mathfrak{S}_p) \rightarrow \Phi(\mathfrak{S}_p)$  is the identity functor in  $MSC_2$ .

Similarly, we have the natural isomorphism  $f: \text{Hom}(\Phi \Psi \mathfrak{L}_q, \mathfrak{L}_q) \rightarrow \text{Hom}(\Psi \mathfrak{L}_q, \Psi \mathfrak{L}_q)$  and therefore we can also calculate the adjunction counit  $\varepsilon: \Psi \circ \Phi \rightarrow id_2$  on components according to the formula:  $\varepsilon_{\mathfrak{L}_k} = f^{-1}(id_{\Psi \mathfrak{L}_k})$  where  $id_{\Psi \mathfrak{L}_k}: \Psi(\mathfrak{L}_k) \rightarrow \Psi(\mathfrak{L}_k)$  is the identity functor in  $MSC_1$ . This function is executed in the  $MSC_2$  model and represents, in our

case, the process of generating the response of the  $MSC_2$  model to the request of the  $MSC_1$  model.

The adjoint of two functors defined by the tuple  $(\Phi, \Psi, \eta, \varepsilon)$ , where  $\Phi$  and  $\Psi$  are adjoint functors,  $\eta$  is the adjoint unit and  $\varepsilon$  is the adjoint counit, determines a monad [11]. An endofunctor  $T: MSC_1 \rightarrow MSC_1$ , together with two natural transformations  $\eta: id \rightarrow T$ , called "return", and  $\mu: T^2 \rightarrow T$ , called "join", which make the diagrams in Fig. 5 and Fig. 6 commutative is a monad in category  $MSC_1$ , which is specified by a tuple  $(T, \eta, \mu)$ .

Based on the adjunction  $(\Phi, \Psi, \eta, \varepsilon)$ , we can calculate the monad components as follows: the endofunctor  $T = \Phi \circ \Psi$ ,  $\eta$  is the adjunct unit, and  $\mu = \Psi \varepsilon \Phi$  [11]. Monads are frequently used in functional languages that offer in this way facilitates for inserting imperative code into the functional code. Therefore, in a similar way, mechanisms for specifying monads can be introduced in domain specific modeling languages [6].

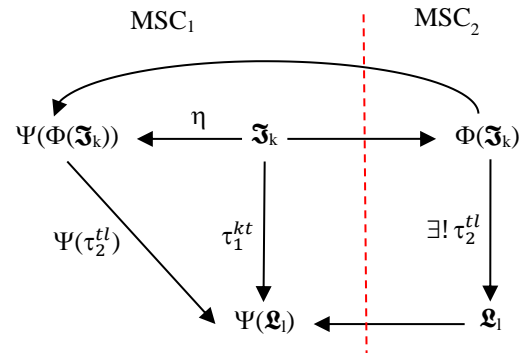


Fig. 3. Definition of Unit Adjunction.

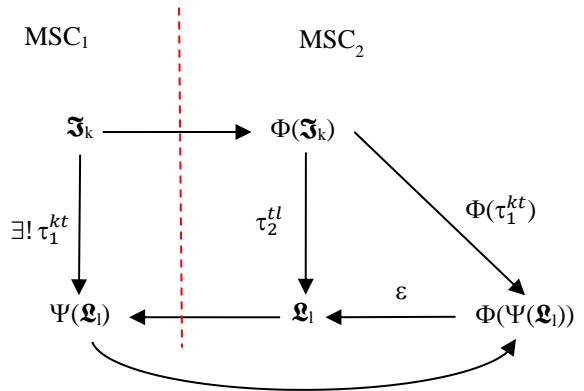


Fig. 4. Definition of Counit Adjunction.

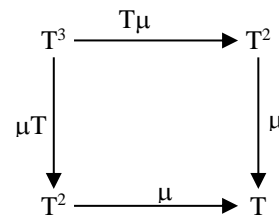


Fig. 5. First Constraint.



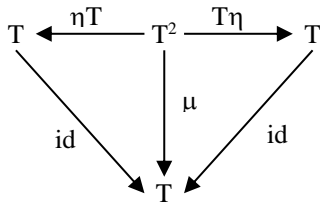


Fig. 6. Second Constraint.

Most of the time, the co-simulation control is realized by an orchestrator that coordinates the system components according to a co-simulation scenario [3,25]. In this case, the two models must be encapsulated in software units that implement a standard interface [2]. The interaction between two CPPS models can be specified by defining shared variables between the two models or by passing data to the orchestrator. In the case of our approach, the adjunction unit  $\eta$  is executed in the  $MSC_1$  model and has the role of preparing the data, in the appropriate format, to be transmitted to the orchestrator. The  $\mu$  application is executed in the  $MSC_1$  model, and specifies the operations to be executed by this model. Connecting the models through monads makes it possible to analyze the composite system resulting from co-simulation.

#### V. ORIGINAL CONTRIBUTIONS AND CONCLUSIONS

The most important finding, in our proposal, is simplicity and conceptual clarity. Thus, the static dimension of a model is the image of a categorical sketch through a functor. The behavioral dimension of a model is specified by a set of functors and a set of behavioral rules. The simulation space of a model is defined by a category. The co-simulation space of two models is specified by a monad induced by two adjoint functors. All the mechanisms involved in these definitions are generic and can be implemented at the metamodel level. Mechanisms for specifying monads can be included in the modeling language at the metamodel level, as happens in functional languages that offer such mechanisms especially to allow imperative specifications. Most of the times the interaction between two CPPS models can be specified by defining shared variables. Connecting the models through monads makes it possible to analyze the composite system resulting from co-simulation. To our knowledge, co-simulation through monads, which is the main objective of this work, has not been addressed until now.

#### REFERENCES

- [1] Henderik A. Proper and Giancarlo Guizzardi, "On Domain Conceptualization" Advances in Enterprise Engineering XIV, EEWK 2020, Bozen-Bolzano, Italy, September 28, October 19, and November 9–10, 2020 ; Springer 2021.
- [2] Functional Mock-up Interface for Model Exchange and Co-Simulation, Document version: 2.0.1 October 2nd 2019, <https://fmi-standard.org/>.
- [3] INTO-CPS Tool Chain User Manual, Deliverable Number: D4.3a Version: 1.0 Date: December, 2017 Public Document, <http://into-cps.au.dk>.
- [4] D. Karagiannis, H.C. Mayr, J. Mylopoulos, "Domain-Specific Conceptual Modeling Concepts, Methods and Tools" Springer International Publishing Switzerland (2016).
- [5] Dominik Bork, Dimitris Karagiannis, Benedikt Pittl, "A survey of modeling language specification techniques", Information Systems 87 (2020) 101425, journal homepage: [www.elsevier.com/locate/is](http://www.elsevier.com/locate/is).

- [6] M. Fowler, R. Parsons, "Domain Specific Languages", 1st ed. Addison-Wesley Longman, Amsterdam, 2010.
- [7] D.C. Crăciunean, D. Karagiannis, "Categorical Modeling Method of Intelligent WorkFlow" in: Groza A., Prasath R. (eds) Mining Intelligence and Knowledge Exploration. MIKE Lecture Notes in Computer Science, vol 11308. Springer, Cham (2018).
- [8] D.C. Crăciunean, "Categorical Grammars for Processes Modeling", International Journal of Advanced Computer Science and Applications(IJACSA), 10(1), (2019).
- [9] D.C. Crăciunean, D. Karagiannis, "A categorical model of process co-simulation", Journal of Advanced Computer Science and Applications(IJACSA), 10(2), (2019).
- [10] Michael Barr And Charles Wells, "Category Theory For Computing Science- Reprints in Theory and Applications of Categories", No. 22, 2012.
- [11] Michael Barr Charles Wells. "Toposes, Triples and Theories" November 2002.
- [12] R. F. C. Walters, "Categories and Computer Science, Cambridge Texts in Computer Science", Edited by D. J. Cooke, Loughborough University, 2006.
- [13] Zinovy Diskin, Tom Maibaum- "Category Theory and Model-Driven Engineering: From Formal Semantics to Design Patterns and Beyond", ACCAT 2012.
- [14] Diskin Z., König H., Lawford M., „Multiple Model Synchronization with Multiary Delta Lenses" in: Russo A., Schürr A. (eds) Fundamental Approaches to Software Engineering. FASE 2018. Lecture Notes in Computer Science, vol 10802. Springer, Cham.
- [15] Uwe Wolter, Zinovy Diskin, "The Next Hundred Diagrammatic Specification Techniques, A Gentle Introduction to Generalized Sketches", 02 September 2015 : <https://www.researchgate.net/publication/253963677>,
- [16] D. Plump, "Computing by graph transformation: 2018/19", Department of Computer Science, University of York, UK, Lecture Slides, 2019.
- [17] G. Campbell, B. Courtehoue and D. Plump, "Linear-time graph algorithms in GP2", Department of Computer Science, University of York, UK, Submitted for publication, 2019. [Online]. Available: <https://cdn.gjcampbell.co.uk/2019/Linear-Time-GP2-Preprint.pdf>.
- [18] D. Plump, "Checking graph-transformation systems for confluence", ECEASST, vol. 26, 2010. DOI: 10.14279/tuj.eceasst.26.367.
- [19] Hartmut Ehrig, Claudia Ermel, Ulrike Golas, Frank Hermann, "Graph and Model Transformation General Framework and Applications", Springer-Verlag Berlin Heidelberg 2015.
- [20] R. Milner, "The Space and Motion of Communicating Agents", Cambridge University Press, (2009).
- [21] C. Gomes, C. Thule, D. Broman, P.G. Larsen, H. Vangheluwe, "Co-simulation: State of the art", ACM Computing Surveys, Vol. 1, No. 1, Article 1. Publication date: January (2016).
- [22] Claudio Gomes, Casper Thule, Levi Lucio, Hans Vangheluwe, and Peter Gorm Larsen, "Generation of Co-simulation Algorithms Subject to Simulator Contracts", <https://sites.google.com/view/cosimcps19>.
- [23] David I. Spivak, "Category Theory for the Sciences", The MIT Press Cambridge, Massachusetts London, England, 2014 Massachusetts Institute of Technology.
- [24] Brendan Fong and David I. Spivak, "Seven Sketches in Compositionality" An Invitation to Applied Category Theory; 2018.
- [25] Bottaccioli, Lorenzo; Estebsari, Abouzar; Pons, Enrico; Bompard, Ettore; Macii, Enrico; Patti, Edoardo; Acquaviva, Andrea, "A Flexible Distributed Infrastructure for Real-Time Co-Simulations in Smart Grids" in: IEEE TRANSACTIONS ON INDUSTRIAL INFORMATICS, vol. 13 n. 6, pp. 3265-3274. - ISSN 1551-3203, (2017).



# A Machine Learning Model for Predicting Heart Disease using Ensemble Methods

Jasjit Singh Samagh, Dilbag Singh  
Department of Computer Science and Engineering  
Chaudhary Devi Lal University, Sirsa, India

**Abstract**—There is the continuous increase in death rate related to cardiac disease across the world. Prediction of the heart disease in advance may help the experts to suggest the preemptive measures to minimize the death risk. The early diagnosis of heart disease symptoms is made possible by machine learning technologies. The existing machine learning models are inefficient in terms of simulation error, accuracy and timing for heart disease prediction. Hence, an efficient approach is needed for efficient prediction of heart disease. In the current research paper, a model based on Machine learning techniques has been proposed for early and accurate prediction of heart disease. The proposed model is based on techniques for feature optimization, feature selection, and ensemble learning. Using WEKA 3.8.3 tool, the feature selection and feature optimisation technique has been applied for irrelevant features elimination and then the pragmatic features are tested using ensemble techniques. Further, the comparison of the proposed model is made with the existing model without feature selection and feature optimisation technique in terms of heart disease prediction effectiveness. It is found that the results of proposed model gives the better performance in terms of simulation error, response time and accuracy in heart disease prediction.

**Keywords**—Heart disease; diagnosis; ensemble; optimization; prediction

## I. INTRODUCTION

Globally, coronary disease is the main cause of death. 15% to 20% of fatalities were thought to be caused by ischemic heart disease and strokes. Investigations including ECGs, chest radiographs, and echocardiograms are typically done at the patient's bedside for the diagnosis of these disorders, although more involved procedures like cardiac catheterization, nuclear scanning, CT scans, and MRIs may also be done. The data that are gathered as a result of these examinations take a long time to analyse, and it takes a long time to deliver medications, which could be harmful to the patient. Doctors, pathologists, and other medical professionals may find that machine learning can shorten test times while improving test accuracy. Data mining and machine learning facilitate quick extraction and helps in fast extraction of results from a large size data in comparison with manual analysis [1].

Data mining plays vital role in health care systems using machine learning. Data mining is regarded as a crucial work that must be carried out precisely and competently since it aims to address real health issues in the diagnosis and treatment of disease. Heart disease prediction model, an electronic approach for detecting the heart illnesses based on

earlier data and information, can be used to determine the sickness and its impact on patients. In order to aid medical professionals in making early forecasts of heart disease, this research aims to replace the time-consuming method with a quick one. Early diagnosis, prompt treatment, and decreased likelihood of casualties [2].

The objective of the present study is to develop and improve ML model which may help medical workers to extend accurate and quick medical help to the needy. For the said purpose different machine learning ensemble algorithms are compared using optimization algorithms. Due to their adaptability, optimization algorithms are frequently employed in numerous research domains. These algorithms are created by simulating or illuminating specific natural processes. GA and PSO are employed for this purpose. GA resolves the given process by replicating the natural process of evaluation. While PSO, a computer method, optimises the problem by iteratively attempting to enhance an optimal solution with regard to a specific number of features. The key features are produced by applying these optimising strategies one after another [10]. The drawbacks of using a single optimization algorithm to solve complicated problems include limited accuracy and generalizability.

In the research, GA and PSO are integrated, which means that exploitation and exploration capabilities are merged for binary and multi-class heart disease in order to further explore the application of optimization in bioinformatics. The wrapper-based feature selection approach is utilised in the study to eliminate redundant and unnecessary features, the GA optimization results are taken into account as the PSO's initial values, and finally the classification model for heart disease is built. Ensemble algorithms including Bagging, Boosting, Randomization, and Hybrid (Integration of Bagging, Boosting, and Randomization) are utilised to conduct the research. For the prediction and classification of the heart disease dataset, it has been found that randomization and hybrid models are better algorithms.

The remaining part of the paper is organized as follows: The problem statement for the research is covered in Section II. The relevant literature is discussed in Section III. The proposed methodology employed for the current study endeavor is described in detail in Section IV. The data mining tool used to conduct the research is discussed in Section V. The dataset for heart disease is presented in Section VI. The experimental findings of the proposed model for the prognosis of heart disease are presented in Section VII. The conclusion and Future Scope is covered in the final part.

## II. PROBLEM STATEMENT

It has been observed in earlier studies that the research is only used to predict and categorise cardiac disease using machine learning approaches. However, the research does not focus on improving these techniques using optimisation techniques, simply on the unique consequences of various machine learning algorithms.

To carry out the work the wrapper-based feature selection method is applied as an initial step, which is also called the pre-treatment step. The mining-relevant attribute selection attributes are chosen from among all the original attributes once all continuous attributes have been discretized. The pre-processing stage of feature selection in machine learning is extremely successful at reducing dimensionality, removing irrelevant data, boosting learning accuracy, and enhancing comprehension of outcomes. The best attributes of the data set are chosen in the following stage by using PSO and GA. The classification will be more accurate thanks to these ideal features. The last phase, which diagnoses heart disease using ensemble approaches, assesses the performance of the suggested procedures by measuring classification accuracy.

## III. LITERATURE REVIEW

Dilbag Singh [11] et al. compared and reviewed already existed models and extracted some major key attributes which are highly useful in creating and building an effective model. The effective model will therefore aid in the early diagnosis of cardiac problems, which will aid medical professionals in the patient monitoring process.

Jasjit Singh Samagh [10] et al. suggested a machine learning model for the heart disease prediction. The model is an integration of wrapper-based feature selection and GA and PSO based feature optimisation technique which is tested on ensemble machine learning techniques. Ensemble techniques itself being the combination of two or more classification techniques resulted in predicting an efficient and more effective model.

Youness Khourdifi [12] et al. associated the algorithms with different performance metrics with the help of machine learning. Different methods were used for prediction. Artificial Neural Networks, Random Forest, and K-Nearest Neighbor provide the greatest outcomes in this investigation. Then the research combines the algorithms and attempted to test the model effectiveness to see if it would be more effective or not. The results were later applied to a data set of heart disease, and his suggested models produced greater accuracy.

K.Vembandasamy [13] et al. analysed some parameters and suggested a data mining-based technique for predicting cardiac disease. Naive Bayes is utilised in the study since it has a strong independence principle for the diagnosis. 500 patients' worth of data from Chennai's premier diabetes research institute were utilised in the study. The tool used WEKA and achieved accuracy of 86.419%.

Vikas Chaurasia [14] et al. suggested the use of data mining tools for heart disease detection. Due to its open source nature and inclusion of machine learning techniques for mining, the author believes WEKA to be the greatest tool. The J48, bagging, and Naive Bayes techniques were applied in the current investigation. The author employed only 11 of the 76 features in the UCI data set, which was used to predict heart disease. Naive is accurately documented with 82.31%. J48 is recorded with an accuracy of 84.35%, and bagging records accuracy at an accuracy of 85.03%.

On the basis of literature review, it can be concluded that to build an effective model for the heart disease prediction it should be an integration of feature selection, feature optimisation and ensemble method.

## IV. PROPOSED METHODOLOGY

The research is carried out in stages for heart disease prediction. The process begins with the collection of data for the dataset. The creation of dataset for the heart disease prediction is the crucial part as ML techniques performance depends upon it. The raw data is collected from electronic source present in form of patient's medical history, observations and laboratory tests. Redundancy is eliminated during pre-processing, which follows data collection, in order to get the information ready for heart disease prediction. The dataset is entered into the data mining tool after it is prepared to assess performance.

The Fig. 1 explains the proposed method for the heart disease prediction. In Proposed method, heart disease training dataset will go through feature selection and feature optimisation technique which will eliminate the low ranked attributes and will result in providing the predominant features. The performance of these predominant features for heart disease prediction is assessed using ensemble approaches [9]. The results predicted by this method are exceptionally best.

### A. Feature Selection

An attribute of a data collection that is used in machine learning is called a feature. Some machine learning experts hold the belief that features should only be applied to attributes that have relevance to a given machine learning problem, although this belief should not be taken at face value. A branch of feature engineering that has attracted a lot of research attention is the selection of the subset of useful features for machine learning. The most important pre-processing step in any machine learning project is probably feature selection. It aims to choose a subset of system information or attributes that contributes most significantly to a machine learning activity [3].

Fig. 2 explains the process of the feature selection that is the internal structure of subset evaluation and resulting in getting the best features.

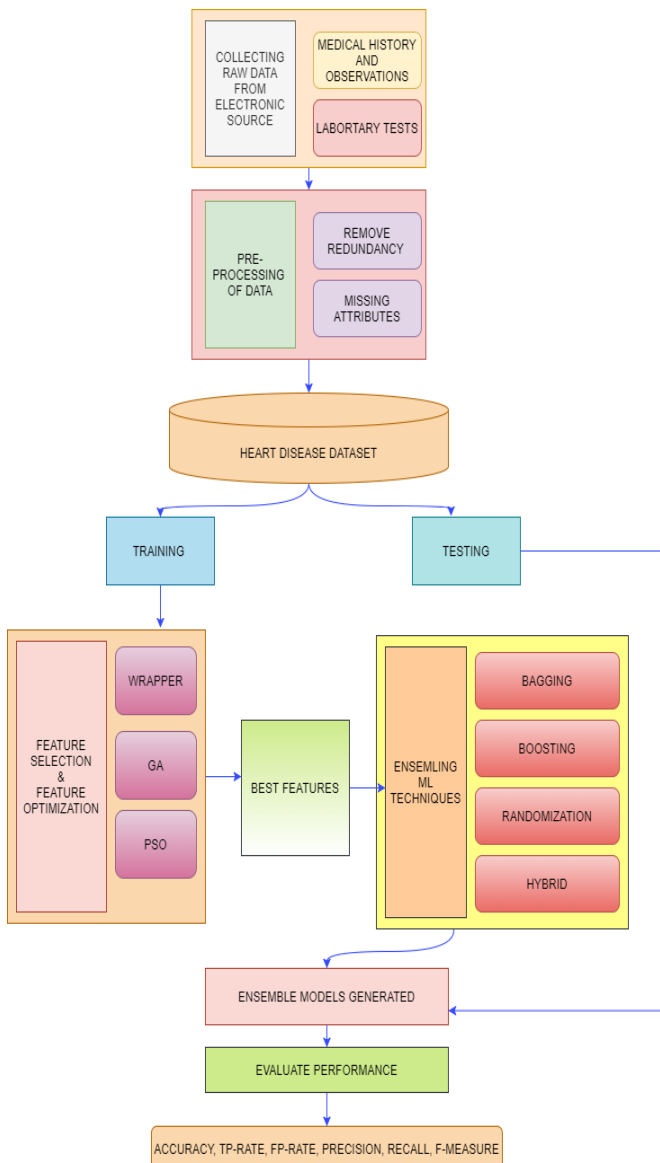


Fig. 1. Proposed Method (with Feature Selection and Feature Optimization).

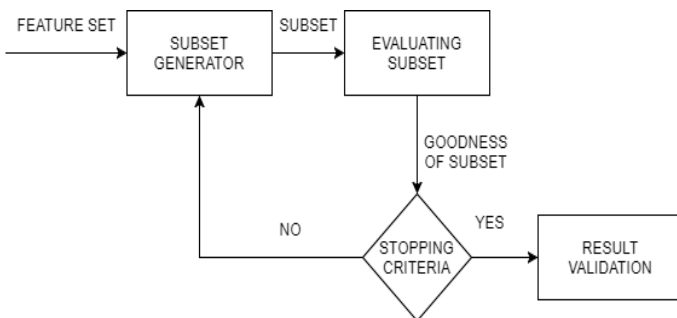


Fig. 2. Process of Feature Selection.

**B. Feature Optimization**

The goal of optimization is to achieve the best outcomes possible in any situation. We use optimization to either maximise the expected benefit or reduce the amount of effort needed. Decision variables can be stated as a function of the

work necessary or the intended benefit. Discovering the circumstances that maximise or reduce a function is, thus, optimization. The highest value of the negative of the function is at point  $x$ , which corresponds to the smallest value of the function  $f(x)$ . As a result, since finding the minimum of a function's negative allows us to find its maximum, optimization can be thought of as a minimization problem. Any optimization problem cannot be solved with a single technique. Consequently, we employ several techniques [4].

- Particle Swarm Optimization (PSO)

Every member of the population is referred to as a particle in particle swarm optimization. Particles in typical PSO update their velocity and location at each iteration based on their own experience, which is their personal best (best), and with the best experience of all the nearby particles, which is the global best (best), after the population has been initialised [7, 16].

A workflow diagram makes it simple to comprehend particle swarm optimization. Fig. 3 depicts this. It demonstrates in detail how PSO operates [8].

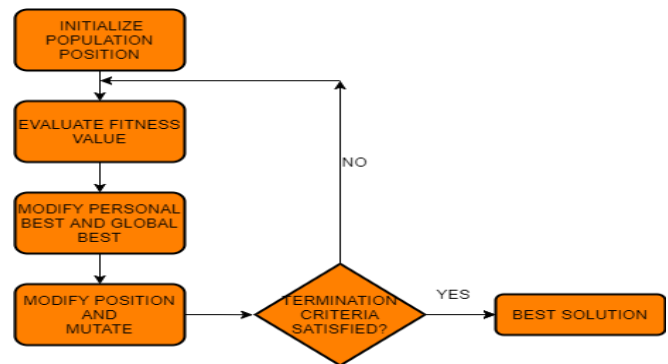


Fig. 3. Workflow of Particle Swarm Optimization (PSO) for Feature Optimization.

- Genetic Algorithm (GA)

The genetic algorithm, an evolutionary computational technique that has gained popularity recently, was created by Holland in the early part of 1975 and later improvised by Goldberg [9]. It is a method of searching that resolves a specific issue by simulating the course of evolution. An algorithm that uses the idea of "survival of the fittest" and is based on Darwin's theory is known as a genetic algorithm [5]. The use of fresh and better optimal solutions by a genetic algorithm is non-presumptive like continuity. The genetic algorithm as a method has enormous potential, and as a result, it has been applied in a variety of industries, including gaming and financial research. The Genetic Algorithm has grown to be highly sought-after in various industries since it can manage a variety of characteristics. Instead of taking a lifetime to solve the problem, it has an ideal solution or one that is very close to the ideal [9].

GA execution can easily be understood with the help of a workflow diagram. Fig. 4 shows the flowchart of GA, and it explains the steps it takes in execution [6].

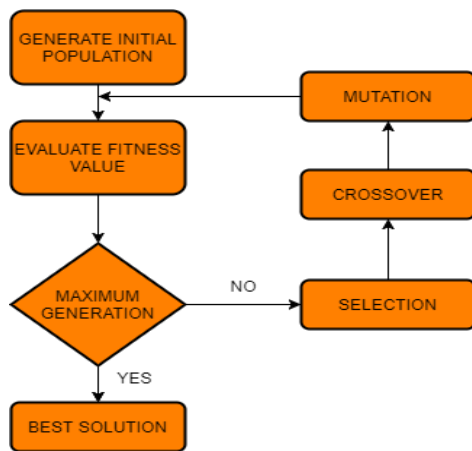


Fig. 4. Workflow of Genetic Algorithm (GA) for Feature Optimization.

### C. Ensemble Techniques

An ensemble is a strategy that combines various models with varied strengths. Ensemble methods combine weaker learners to create stronger ones. Ensemble learning models are used to improve the predictive performance of statistical learning techniques by constructing a linear mixture of the simpler base learner. Ensemble learning models use decision trees as the base learner; in the case of random forest, many boosting and bagging implementations have been proposed [17].

One of the earliest and the most popular ensemble models is bootstrap aggregating or bagging. Bagging uses bootstrapping to generate multiple training datasets, and utilizing the same learning process, a collection of models are created using these training datasets [15].

Boosting is another crucial ensemble-based approach, similar to bagging. With boosting, weaker learning models are trained on resampled data, and the results are blended based on the performance of many models and a weighted voting mechanism. A specific type of boosting algorithm is adaptive boosting, often known as AdaBoost. Another well-liked ensemble learning strategy for creating prediction models is randomization [15, 17].

### V. DATA MINING TOOL

When doing numerous experiments on datasets for machine learning, the data mining tool is crucial. WEKA 3.8.3 is the data mining tool utilised to carry out the current study. Waikato Environment for Knowledge Analysis is abbreviated as WEKA. It is created in the New Zealand's University of Waikato. Its platform is Java-based [21].

WEKA is a collection of AI algorithms used for data mining jobs. There are two ways to apply an algorithm when using WEKA, either directly on a dataset or by calling it from Java code. Numerous built-in features in WEKA allow for the prediction of the model's correctness [19].

### VI. HEART DISEASE DATASET

Data could be gathered for the research's objective from a variety of sources. The UCI machine learning repository and Kaggle are the two most frequently used sources. In 1987 at

Irvine, David Aha and a few other students founded the UCI repository [22], whereas the Kaggle is Google Subsidiary founded by Anthony Goldbloom in April, 2010 at United States.

Both UCI machine learning repository [18] and Kaggle contains number of datasets related to healthcare sector. Machine learning enthusiasts, from novices to experts, frequently use the dataset available from these sources to comprehend data empirically.

Further, to carry out the present research the dataset from Staglog (heart) dataset from UCI repository is used which is provided by university hospital, Basel, Switzerland by Matthias Ptisterer, M.D. The data is then pre-processed to check for missing attributes and to eliminate duplication. Machine learning is used to forecast heart disease using a later dataset. The heart disease dataset utilised in the study has features that can number in the tens of thousands, but it also contains 14 attributes that are listed below [20].

TABLE I. HEART DISEASE DATASET

Attributes	Types	Explanation
Age	Numeric	Age in years (29 to 77)
Sex	Numeric	Sex (1 = Male, 0 =Female)
Chest Pain Type (cp)	Numeric	Chest pain type (1: typical angina, 2: atypical angina, 3: non-angina pain, 4: asymptomatic)
Rest Blood Pressure (restbtps)	Numeric	Resting blood pressure ( in mm Hg on admission to the hospital) [94, 200]
Serum Cholesterol (chol)	Numeric	Serum cholesterol in mg/dl [126, 564]
Fasting Blood Sugar (fbs)	Numeric	Fasting blood sugar > 120 mg/dl (1 = True, 0= False)
Resting Electrocardiographic (restecg)	Numeric	Resting Electrocardiographic(ECG) results values(0: normal, 1: having ST-T wave abnormality, 2:showing probable left ventricular hypertrophy)
Maximum Heart Rate (thalach)	Numeric	Maximum heart rate achieved [71, 202]
Exercise Induced (exang)	Numeric	Exercise induced angina (1: Yes, 0: No)
Oldpeak	Numeric	ST depression induced by exercise relative to rest [0.0,62.0]
Slope	Numeric	The slope of the peak exercise ST segment (1: upsloping, 2: flat, 3: downsloping)
Major Vessels (ca)	Numeric	Number of major vessels values(0-3) colored by flourosopy
Thal	Numeric	Defect types: value 3: normal, 6: fixed defect, 7: irreversible defect
Class	Nominal	Prediction of heart disease (1: Absence, 2: Presence)

Table I is the Heart disease dataset used in predicting the heart disease. It consists of the 14 attributes with its explanation used for predicting the model.

**A. Feature Selection Feature Optimisation and Ensembling**

The databases for cardiac disease can contain up to tens of thousands of features. However, only roughly 14 qualities are necessary to forecast cardiac problems because a large number of irrelevant and redundant attributes frequently produce unreliable results, incur expensive costs, and take a lot of time. The performance in predicting heart disease would be better if the qualities were less. On the other hand, Feature Optimization techniques [25] used to identify the optimal solution by minimising or maximising the objective function without going against resource limitations.

By adjusting parameters, optimization methods can be used to enhance the performance of classifiers in the prediction of heart disease. GA and PSO are employed for this purpose [8]. While PSO, a computer method, optimises the problem by iteratively attempting to enhance a candidate solution with regard to a specific number of features, GA resolves the given process by replicating the natural process of evaluation. These optimization strategies are repeatedly used to produce the key features [17].

After applying the wrapper, Genetic Algorithm and Particle swarm optimisation to the heart disease dataset following features got extracted:

- Chest Pain Type (cp).
- Resting Electrocardiographic (restecg)
- Maximum Heart Rate (thalach).
- Exercise Induced (exang).
- Oldpeak.
- Major Vessels(ca).
- Thal.
- Class.

The number of features reduced from 14 attributes to the 8 predominant features. These features are then tested on ensemble techniques that is Bagging, Boosting, Random Forest and Hybrid (a combination of Bagging, Boosting and Randomization) methods of machine learning to evaluate the results regarding prediction of heart disease.

**VII. EXPERIMENTAL RESULTS**

To evaluate the performance, comparison of various ensemble machine learning techniques is made on various criteria.

The Results are calculated on the basis of following norms:

- Results without using Feature selection and Feature Optimization Techniques.
- Results after using Feature Selection and Feature Optimization Techniques (Proposed method).

**A. Simulation Error of Ensemble's**

Simulation error is also taken into consideration in the study's execution to enhance the ensemble learning model's performance. The prediction model's effectiveness is described by simulation errors. The five aforementioned evaluation criteria are used in the current study to assess the simulation error, as indicated in Table II and III, respectively.

Table II describes the simulation error for the heart disease prediction model using ensemble technique. Simulation error tell us how effective the model is in predicting the accuracy. In this, result evaluated is computed without using feature selection and feature optimisation technique.

Table III is the simulation error table for the heart disease prediction model using ensemble technique. In this table the results are computed using feature selection and feature optimisation technique. It is seen that the error rate of the simulation model are reduced after using feature selection and feature optimisation technique that is proposed method.

**B. Confusion Matrix**

The confusion matrix reveals how a predictive model operates internally. It provides insight information on classes that are accurately predicted, classes that are wrongly forecasted, and also about the different forms of faults. The confusion matrix produced for a two-class classification issue with negative and positive classes is the most basic type of confusion matrix. Each cell in the figure has a distinct and clear display, as illustrated. [23].

Fig. 5 explains the insight of the confusion matrix, that how the internal functional of the predictive model are classified in following classes.

TABLE II. SIMULATION ERROR WITHOUT USING FEATURE SELECTION AND FEATURE OPTIMISATION

Evaluation Criteria	Bagging	Boosting	Randomization	Hybrid
KS	0.58	0.595	0.6244	0.6694
MAE	0.2934	0.2374	0.2696	0.2668
RMSE	0.3774	0.3807	0.3587	0.362
RAE (%)	59.4167 %	48.0607 %	54.5794 %	54.0189 %
RRSE (%)	75.956 %	76.6224 %	72.1837 %	72.8601 %

TABLE III. SIMULATION ERROR WITH FEATURE SELECTION AND FEATURE OPTIMISATION

Evaluation Criteria	Bagging	Boosting	Randomization	Hybrid
KS	0.7657	0.6772	1	0.7963
MAE	0.2271	0.2121	0.0867	0.1753
RMSE	0.3001	0.3296	0.1346	0.2474
RAE (%)	45.9852 %	42.9527 %	17.5484 %	35.4954 %
RRSE (%)	60.3894 %	66.3318 %	27.0926 %	49.7841 %

Where KS- Kappa Statistic, MAE- Mean Absolute Error, RMSE- Root Mean Squared Error, RAE- Absolute Error and RRSE- Root Relative Squared Error.

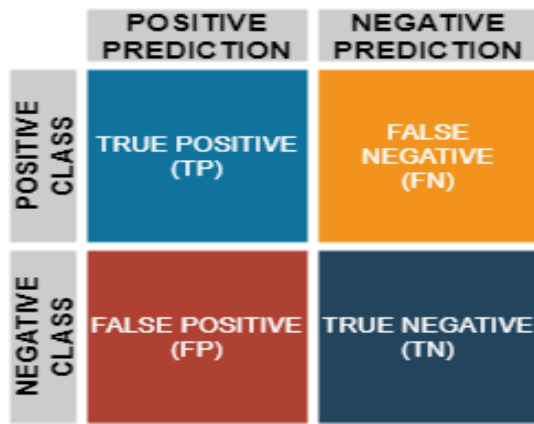


Fig. 5. Insight of Confusion Matrix.

- True Positive (TP): When both the truth and the test's prediction of a positive are true, the class is true positive. For instance, when a patient is ill and the test also detects this.
- True Negative (TN): When the test also predicts a negative outcome and the truth is negative, the class is true negative. For instance, when a test accurately detects that a person is healthy.
- False Negative (FN): When the truth is positive but the test predicts a negative outcome, the class is considered false negative. For instance, when a test falsely indicates that someone is healthy when they are actually ill. In statistics, this is also known as Type II mistake.
- False Positive (FP): The term "false positive" refers to a situation in which the test anticipates a positive result even though the reality is different. If a test falsely indicates that a person is ill even when they are not ill. In statistics, this is known as Type I mistake. [23, 24].

The Table IV indicates the Confusion Matrix of the Ensemble Machine learning models for prediction of the heart disease. The table shows detailed confusion matrix of machine learning ensemble model when created without using feature selection and feature optimisation technique that is the traditional method the proposed method that is using feature selection and feature optimisation techniques. This matrix depends upon four factors namely TP, TN, FP and FN [23, 24]

C. Accuracy Factors

The factors on which the accuracy of model is based are TP rate, FP rate, Precision, Recall and F measures [24].

Fig. 6 is about the calculations of accuracy factors. It shows the formulas used to calculate the accuracy factors Such as TP rate, FP rate, Precision, Recall and F-measure.

Following model creation, its accuracy is assessed by contrasting it against the following criteria.

- True positive rate: that a real positive will test positive is known as the true positive rate. It is calculated as TP/TP+FN and is often referred to as sensitivity.

- Precision: is the ratio of examples that genuinely belong to a class to all instances that are categorised in that class.
- Recall: a class's true total is equal to its actual fraction of instances classed as that class (equivalent to TP rate).
- F-Measure: is a composite measure for recall and precision that is calculated as 2 \* recall / (precision + recall).

TABLE IV. CONFUSION MATRIX OF MODELS

		Absent	Present	Class
Without Feature Selection and Feature Optimisation	Bagging	122	28	Absent
		28	92	Present
	Boosting	123	27	Absent
		27	93	Present
	Randomization	126	24	Absent
		24	94	Present
	Hybrid	129	21	Absent
		23	97	Present
With Feature Selection and Feature Optimisation	Bagging	139	11	Absent
		20	100	Present
	Boosting	129	21	Absent
		22	98	Present
	Randomization	150	0	Absent
		0	120	Present
	Hybrid	140	10	Absent
		17	103	Present

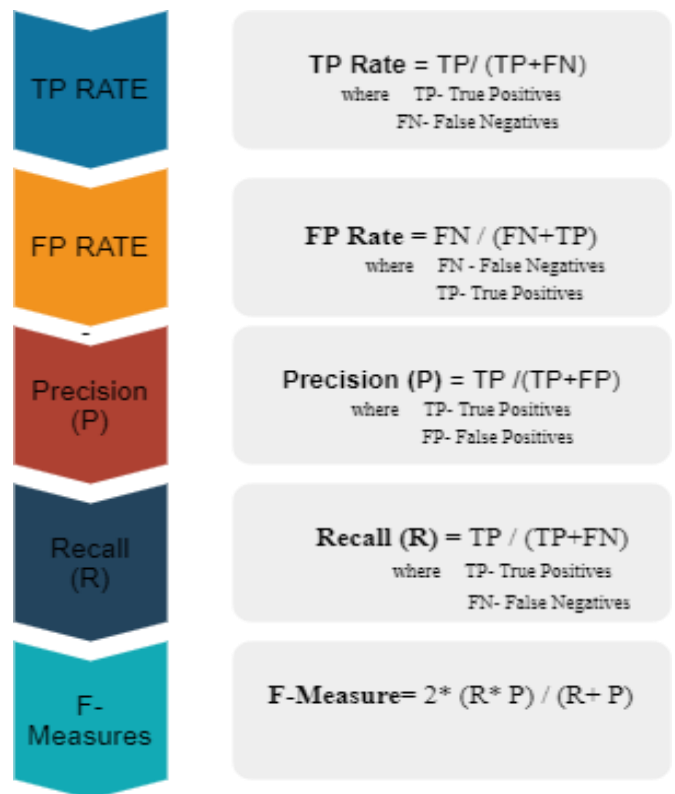


Fig. 6. Calculations of Accuracy Factors.



TABLE V. DETAILED ACCURACY MEASURES OF FACTORS WITH CLASS

		TP Rate	FP Rate	Precision	Recall	F-Measure	Class
Without Feature Selection and Feature Optimization	Bagging	0.813	0.233	0.813	0.813	0.813	Absent
		0.767	0.187	0.767	0.767	0.767	Present
	Boosting	0.820	0.225	0.820	0.820	0.820	Absent
		0.775	0.180	0.775	0.775	0.775	Present
	Randomization	0.840	0.217	0.829	0.840	0.834	Absent
		0.783	0.160	0.797	0.783	0.790	Present
	Hybrid	0.860	0.192	0.849	0.860	0.854	Absent
		0.808	0.140	0.822	0.808	0.815	Present
With Feature Selection and Feature Optimization	Bagging	0.927	0.167	0.874	0.927	0.900	Absent
		0.833	0.073	0.901	0.833	0.866	Present
	Boosting	0.860	0.183	0.854	0.860	0.857	Absent
		0.817	0.140	0.824	0.817	0.820	Present
	Randomization	1.000	0.000	1.000	1.000	1.000	Absent
		1.000	0.000	1.000	1.000	1.000	Present
	Hybrid	0.933	0.142	0.892	0.933	0.912	Absent
		0.858	0.067	0.912	0.858	0.884	Present

The Table V indicates the Accuracy factors of the Ensemble Machine learning models for prediction of the heart disease. The table shows detailed accuracy measures of factors with class of machine learning ensemble model when created without using feature selection and feature optimisation technique, that is, the traditional method (the proposed method) to use feature selection and feature optimisation techniques.

depicted from the line chart that the model with feature selection and feature optimisation techniques have better results than the model without feature selection and feature optimisation technique. Moreover, an increase in accuracy in all the stated ensemble techniques, after the use of feature selection and feature optimization, is observed.

### VIII. CONCLUSION AND FUTURE SCOPE

One of the leading causes of death in the modern world is heart disease. However, if it can be predicted beforehand, it can give clinicians crucial information for diagnosis and treatment. In order to prevent cardiac diseases, it is essential to keep track of any health issues. The effectiveness of machine learning in making these predictions has been demonstrated, and the dataset can be used to derive some interesting conclusions. Though there are many existing machine learning models, whose performance needs to be improved in terms of accuracy, consumes more time and are more prone to simulation errors. The prediction behaviour of the model depends upon these factors. A machine learning model built on ensemble learning and using feature selection and feature optimization techniques is suggested to reach this goal.

The experimental findings for the prediction of heart disease utilising ensemble machine learning approaches are presented in this work. To carry out the current study, exploratory, experimental, and applied research approaches were employed. Data about the patient is gathered from the UCI repository. Experiments are conducted using heart disease dataset which are applied on Weka3.8.3 a data mining tool to predict results. Experiments are conducted on two methods, one without using feature selection and feature optimisation method and another on proposed method, using feature selection and feature optimisation technique and later these methods are tested on Ensemble techniques to evaluate the results.

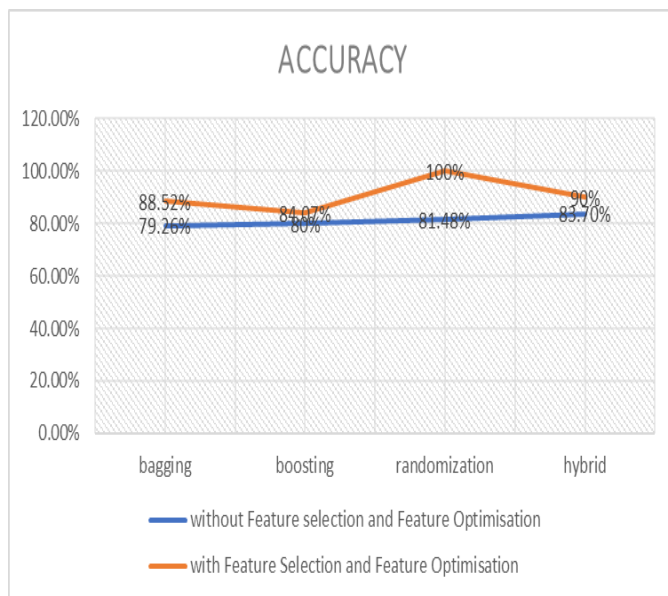


Fig. 7. Comparison of Performance of the Models with and without Feature Selection and Feature Optimisation.

Fig. 7 illustrates the comparison of performance of the model with and without feature selection and feature optimisation technique in heart disease prediction. It is

It is seen that results are better with the proposed technique. Thus, the method can be utilised by medical professionals to forecast and detect cardiac illness early on, which helps to avoid problems.

In future, the same approach might be applied to various datasets of various diseases gathered from various sources.

#### REFERENCES

- [1] John A.A. Hunter, Editor, "Davidson's Principles and Practice of Medicine", Churchill Livingstone An Imprint of Elsevier Limited, (2002).
- [2] Subha, R. (2016). Study On Cardiovascular Disease Classification Using Machine Learning Approaches. *International Journal Of Engineering And Computer Science*. doi: 10.18535/ijecs/v4i12.16.
- [3] Alabdulwahab, S., & Moon, B. (2020). Feature Selection Methods Simultaneously Improve the Detection Accuracy and Model Building Time of Machine Learning Classifiers. *Symmetry*, 12(9), 1424. doi: 10.3390/sym12091424.
- [4] B. Dengiz, F. Altiparmak and A. E. Smith, Local search genetic algorithm for optimal design of reliable networks, *IEEE Trans. Evol. Comput.* 1 (1997).
- [5] G. R. Harik, F. G. Lobo and D. E. Goldberg, *IEEE Trans. Evol. Comput.* 3 (1999), 287–297.
- [6] C. Miles, S. J. Louis, N. Cole and J. McDonnell, Learning to play like a human: case injected genetic algorithms for strategic computer gaming, in: *Proc. 2004 Congr. Evol. Comput.* (IEEE Cat. No. 04TH8753), vol. 2, pp. 1441–1448, IEEE, Portland, OR, USA, 2004.
- [7] Indu Yekkala, Sunanda Dixit and M.A. Jabbar, "Prediction of heart disease using Ensemble Learning and Particle Swarm Optimization", *International Conference on Smart Technology for Smart Nation*, (2017).
- [8] Russell Eberhart and James Kennedy, "A New Optimizer Using Particle Swarm Theory", *IEEE* (1995), pp.39-43.
- [9] Twardowski, K. (1994). An associative architecture for genetic algorithm-based machine learning. *Computer*, 27(11), 27-38. doi: 10.1109/2.330041.
- [10] Jasjit Singh Samagh, Dilbag Singh, "Machine Learning Based Hybrid Model for Heart Disease Prediction" *Annals of the Romanian Society for Cell Biology*, 2199, (2021).
- [11] Dilbag Singh and Jasjit Singh Samagh, "A Comprehensive Review of Heart Disease Prediction using Machine Learning", *Journal of Critical Reviews*, vol. 7, no. 12, (2020), pp.281-285.
- [12] Youness Khourdifi and Mohamed Bahaji, "Heart Disease Prediction and Classification Using Machine Learning Algorithms Optimized by Particle Swarm Optimization and Ant Colony Optimization", *International Journal of Intelligent Engineering and Systems*, Vol. 12, no. 1, (2019).
- [13] K.Vembandasamy, R.Sasipriya and E.Deepa, "Heart Diseases Detection Using Naive Bayes Algorithm", *International Journal of Innovative Science, Engineering & Technology*, Vol. 2 Issue 9, (2015).
- [14] Vikas Chaurasia and Saurabh Pal, "Data Mining Approach to Detect Heart Diseases", *International Journal of Advanced Computer Science and Information Technology (IJACSIT)* Vol. 2, No. 4, Month Year, Page: 56-66, ISSN: 2296-1739, (2012).
- [15] Soni, A. (2020). Diabetes Mellitus Prediction Using Ensemble Machine Learning Techniques. *SSRN Electronic Journal*. doi: 10.2139/ssrn.3642877.
- [16] J. Kennedy, "Particle Swarm Optimization", in *Encyclopedia of Machine Learning* (Springer), C. Sammut and G. I. Webb, Eds. Boston, MA: Springer US, pp. 760–766, (2010).
- [17] Saikat Dutt, Subramanian Chandramouli and Amit Kumar Das, "Machine Learning", Pearson India Education Services Pvt. Ltd, India, (2019).
- [18] M. Lichman, "UCI Machine Learning Repository", [Online]. <https://archive.ics.uci.edu/>, 2013.
- [19] Fatma Zahra Abdeldjouad, Menaouer Brahami, and Nada Matta, "A Hybrid Approach for Heart Disease Diagnosis and Prediction Using Machine Learning Techniques", *ICOST 2020, LNCS 12157*, pp. 299–306, (2020).
- [20] U. H. Dataset, "UCI Machine Learning Repository", [online]. [https://archive.ics.uci.edu/ml/datasets/statlog+\(heart\)](https://archive.ics.uci.edu/ml/datasets/statlog+(heart)).
- [21] Eibe Frank, Mark A. Hall, and Ian H. Witten (2016). *The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques"*, Morgan Kaufmann, Fourth Edition, 2016.
- [22] Dua, D. and Graff, C. (2019). *UCI Machine Learning Repository* [<http://archive.ics.uci.edu/ml>]. Irvine, CA: University of California, School of Information and Computer Science.
- [23] Tyagi, N., 2022. What is Confusion Matrix? | Analytics Steps. [online] [Analyticssteps.com](https://www.analyticssteps.com/blogs/what-confusion-matrix). Available at: <<https://www.analyticssteps.com/blogs/what-confusion-matrix>>.
- [24] Brownlee, J. (2022). How to Calculate Precision, Recall, and F-Measure for Imbalanced Classification. Retrieved from <https://machinelearningmastery.com/precision-recall-and-f-measure-for-imbalanced-classification/>.
- [25] H. Ceylan and M. G. H. Bell, Traffic signal timing optimisation based on genetic algorithm approach, including drivers' routing, *Transport. Res.* 38 (2004), 329–342.

# Novel Approach in Classification and Prediction of COVID-19 from Radiograph Images using CNN

Chalapathiraju Kanumuri<sup>1</sup>  
Department of ECE  
S.R.K.R Engineering College  
Bhimavaram, India

CH. Renu Madhavi<sup>2</sup>  
Department of EIE  
RV College of Engineering  
Bangalore, India

Torthi Ravichandra<sup>3</sup>  
Department of ECE  
Ellenki College of Engineering  
Hyderabad, India

**Abstract**—Effective screening and early detection of COVID-19 patients are highly crucial to slow down and stop the disease's rapid spread at this time. Currently, RT-PCR, CT scanning and Chest X-ray (CXR) imaging are the diagnosis mechanisms for COVID-19 detection. In this proposed work radiology examination by using CXR images is used for COVID-19 detection due to dearth of CT Scanners and RT-PCR testing centers. Therefore, researchers have developed various Deep and Machine Learning systems that can predict COVID-19 using CXR images. Out of which, few are exhibited good prediction results. However, Most of the models are suffered with over fitting, high variance, memory and generalization errors which are caused by noise as well as datasets are limited. Therefore, a Convolutional Neural Network (CNN) with the leveraging Efficient Net architecture is proposed for COVID-19 case detection. The proposed methods have an accuracy of 99% which gives the better results than the present available methods. Therefore, the proposed model can be used in real-time covid-19 classification systems.

**Keywords**—COVID-19; x-ray images; deep learning technique; CNN; efficient net

## I. INTRODUCTION

SARS-CoV-2 has never been discovered in people before December 2019, is the virus that causes the unique Coronavirus Disease 2019 (COVID-19), an infectious and lethal disease that has never been seen before in the world [1]. Finding infected individuals through efficient screening is a crucial responsibility in halting the rapid spread of COVID-19 so that they can be separated and given prompt medical attention. The RT-PCR test is now the most widely utilised screening method for COVID-19 case detection [2]. There are still a number of difficulties with RT-PCR testing, despite it being acknowledged as the "gold standard" for identifying infected cases of the disease. Tahamtan et al recent study [3] found that the detection sensitivity is highly varied and can lead to both FN and FP i.e. False-Negative & False-Positive results. The Radiograph imaging and CT Scan are performed and analysed by radiologists to determine whether or not a suspected person was Infected by Covid-19. This alternative effective screening method to RT-PCR is for the quick identification of COVID-19. All three of these tests are typically used to diagnose COVID-19. The RTPCR test, which is used to diagnose viral infection, can identify viral RNA in sputum or nasopharyngeal swabs [4]. Using CT scans, a 3D radiographic image is examined from various perspectives as part of a CT-based examination. The main drawback in the CT

scan is that, it requires lots of time and the equipment is not readily available all the time and in all the areas and high radiation. Although CT scans are more sensitive to pulmonary disorders, they have a number of drawbacks that prevent them from being used in COVID-19 case detection on a broader scale. These drawbacks include non-portability, prolonged scanning, and the potential for patient exposure. In compared to CT scans, CXR imaging offers an adequate level of accuracy in the detection of COVID-19 cases while being portable, quicker, and more widely accessible and less expensive. Due to these advantages, CXR image analysis for COVID19 case detection has become the focus of numerous recent investigations [5, 6]. With the pandemic's rapid spread, certain studies, in particular, advise using portable CXR imaging as a reliable tool for finding COVID-19 cases.

Many writers recommended combining the RT-PCR test with additional clinical procedures like the CT & CXR (chest X-ray). A few recent studies provide estimates of the sensitivity of professional radiologists to diagnose COVID-19 using CT scans, RT-PCR, and CXR. A research on 51 individuals who had a chest CT and RT-PCR performed within three days found that the CT had a sensitivity of 98% and the RT-PCR had a sensitivity of 71%. Similarly RT-PCR will require at least 12Hrs of time, which is not ideal as COVID19 +ve patients should be identified and tracked as soon as feasible, and it requires specialised materials and equipment that are not readily available [7][8].

A sensitivity of 69% for CXR was found in a different study on 64 patients. An analysis of 636 ambulatory patients revealed that the majority of patients with confirmed coronavirus who visit urgent care facilities have normal or barely abnormal CXR results. The professional eye accurately diagnoses only 58.3% of these cases. Although CXR imaging is fairly quick, COVID-19 case detection must be done manually by qualified radiologists, which takes professional knowledge and is a laborious process. However, there are many fewer radiologists than there are people who are being monitored. These scenarios will all broaden the application of AI-driven algorithms for detecting COVID-19 from chest radiographs.

## II. PREVIOUS WORK

Thus, a diagnostic system powered by artificial intelligence (AI) is required to help radiologists screen COVID-19 cases in a more quick and reliable manner. Without such a system, it is

likely that infected individuals may not be promptly identified, isolated, and treated. Based on the AI[17] aided diagnosis many authors proposed different approaches in diagnosing Covid-19 which are based on CT-scan and radiographs along with the technology end depends on transfer learning and machine learning etc.

CNN based data driven deep learning methods have shown promising performance for the classification challenge of COVID-19 case detection with CXR pictures, which is essentially a machine learning problem. As a result, there are numerous recent research that seek to train new deep learning models for infected case detection with CXR images by reusing or changing existing deep neural networks on top of gathered CXR image datasets. Although several research in their articles indicate considerably higher prediction accuracy for their proposed deep learning models with their own datasets, in practice, noise and restricted training data size may cause a deep learning model to suffer from over fitting, excessive variance, and generalization mistakes.

Deep learning (DL) methods for automated image processing have the potential to significantly improve the role of CXR pictures in the rapid diagnosis of COVID-19[20]. A reliable and accurate DL model could enhance medical decision-making and be used as a triage tool. Recent investigations claim to have achieved outstanding sensitivities > 95%, which is much higher than expert radiologists. Numerous deep learning & Transfer Learning based AI-assisted detection techniques have been proposed to lessen the load of detection from radiography pictures (e.g., CT and CXR images) for radiologists [09][10].

CXR pictures are evolving into a well-liked and often used data source for COVID-19 case detection due to its many advantages over CT images, including mobility, availability, accessibility, and quick testing. When AlexNet, ResNet18, DenseNet201, and SqueezeNet were used to identify two classes (i.e., COVID-19 and Normal) for CXR images, Wynants, L. et al. [11] came to the conclusion that SqueezeNet performed better than the other neural networks. Instead, Narin et al.'s [18] comparison of various CNN models trained on CXR images for COVID-19 case detection, including ResNet-50, Inception V3, and Inception-ResNetV2, revealed that ResNet-50 surpasses the other two models with 98 percent accuracy.

A three class prediction with the help of transfer learning was proposed by E. H. Chowdhury et al. and attained an accuracy of 99% by implementing it in the CNN environment [12]. Farooq et al. [13] provided a COVID-ResNet by fine tuning a pre-trained ResNet-50 architecture with a reported accuracy of 96.23%. A semi-supervised few-shot segmentation for the detection of Covid-19 was proposed by Mohamed Abdel-Basset et al. [14] which is a Deep learning based architecture and the implementation was based on CT scans, the limitation of less data set availability overcomes by this approach. In the similar manner with the DL and TL approaches in the diagnosis of COVID-19 from the X-ray and CT images have been implemented in Halgurd S. Maghdid et al. approach [15]. Linda Wang et al. [16] came up with a huge chest X-ray dataset based approach by a DCNN design and

attained good accuracy with the model with this huge dataset system can encounter technical issues [19].

### III. PROPOSED METHOD

#### A. Datasets

The dataset of 15000+ of (512x512pixel) images are collected from Kaggle which are used to train and test the model. The dataset consists of properly labeled radiograph images by radiologists. When a lot of patients need to be examined in a short amount of time, a model like this can help medical practitioners diagnose COVID-19 cases considerably faster than a radiologist having to go over each scan one by one. The sample images are shown in Fig. 1.

#### B. Learning Rate

Learning rate plays an important role in defining the accuracy of the model. In general convergence can be improved by Low Learning rate but the accuracy will be not as expected. Similarly overshooting can be possible with higher learning rate, therefore choosing learning rate properly will improve the performance there by accuracy. In proposed method the model experience with  $10e-4$  rate which led the model with low error rates on test dataset.

#### C. Batch Size

With reference to the literature article [22] batch size in association with learning rate plays a vital role in accuracy. Increase in batch size like 256 or 128 or 64 might lead the system into memory issues. Therefore depends on system performance and other factors in proposed method tested with 32. Hence proposed model trained with a batch size of 32 and learning rate of  $10e-4$ .

#### D. Epochs

The EfficientNet b3 model consists of 12 Million parameters and the images used in the proposed model have chosen for the training instance is around 15k+ images. With the dataset trained model attained accuracy of 99% at 7th epoch. And the system used about 10 epochs with different batch sizes and learning rate combinations.

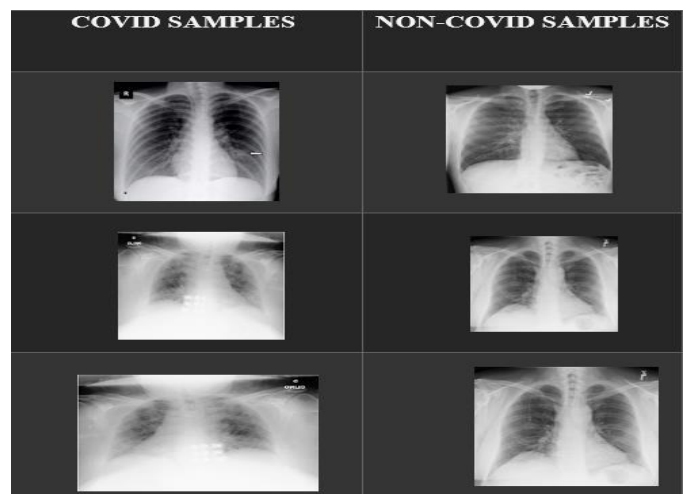


Fig. 1. Images taken as Sample for Covid & Non-Covid.

E. Pre-Trained Model

In the current quickly evolving technology era, applying this technology to the medical diagnosis will leads to better outcome. Computer Vision is one of the fields where we make a use of diagnosing Covid-19. With reference to the Google AI Blog [21] in our proposal we have chosen EfficientNet in which b3 version has chosen to overcome issues in system configuration instead of using VGG16 or Resnet50. In general most of the architectures are too wide, deep or with high resolution therefore while increasing these characteristics will initially help the model but later it leads to saturation and there by become in-efficient, where as in the chosen EfficientNet model these are scaled in a more principled way [23].

Coming to the architecture representation of EfficientNet-B3 will be shown in Fig. 3.

The difference that we can observe between different EfficientNet models will be the number of feature maps i.e. channels, which leads to increase in number of parameters.

The proposed model initially trained on Google Colabs environment which is an effective tool for learning and quickly creating machine learning models in python with pytorch. Later, implemented the same model in the system environment (Fig. 2).

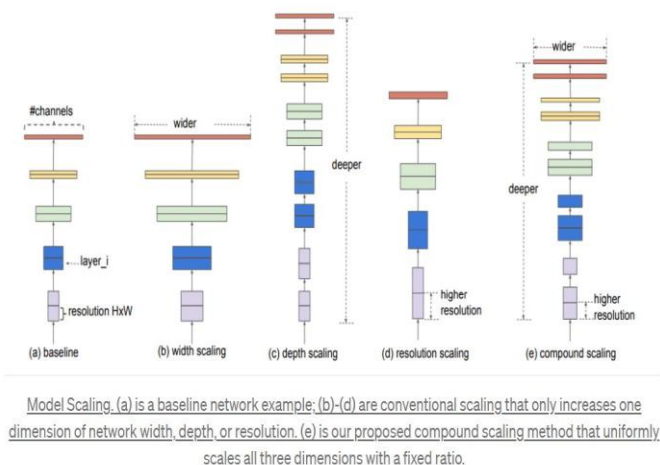


Fig. 2. Model Scaling.

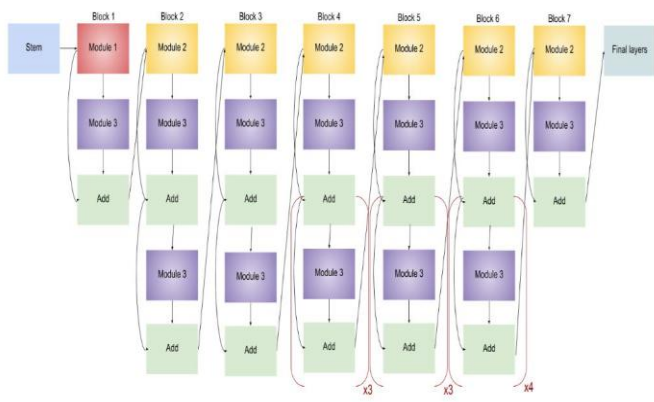


Fig. 3. Architecture for EfficientNet-B3.

Development of the proposed work implemented with the help of PyTorch which will be well suited for the Computer vision applications. When we give radiograph image as input the trained model has to detect the provided image was COVID-19 positive or Negative.

F. Flowchart

The proposed model flow of implementation can be seen in Fig. 4. The existing Dataset consist of the images for training and testing. By applying the transformation method to resize the images and this transformation will be helpful in data augmentation. If more than 90% of the photos in the training dataset belong to the non-Covid class, the training dataset is markedly unbalanced. A naive learning method that just outputs the class of the majority class as the output would achieve high accuracy with severely unbalanced sets, which is a problem. In other words, even though only 10% of people truly have Covid, it will classify everyone as being Covid-free and achieve a "accuracy" of 90%. For the training images with the help of this transform method we can up sample the minority class which is Covid, so that in both the classes we can have same number of images which will be quite helpful improving the accuracy of the model. Now these newly created images can be split into training and validation at a ratio of 80:20. These images are given as input to the system. Across the preprocessing with the help of Data loader Batch size of 32 applied on the training dataset and shuffled them to ensure approximately equal representation of both classes in all the batches. While this process is taking place it's always best practice to save the weights generated at the time of training process to ensure that even though the model got crashed due to any technical reason, instead of starting from the scratch we can call the saved weights and reinitiate the process. Initially while doing the training in the Google colabs the model got crashed because of the number of training images are huge in number and it cannot support these many iterations on the training data as well as testing. Therefore saving the weights will always the best way. And that's why developed model implemented on system environment.

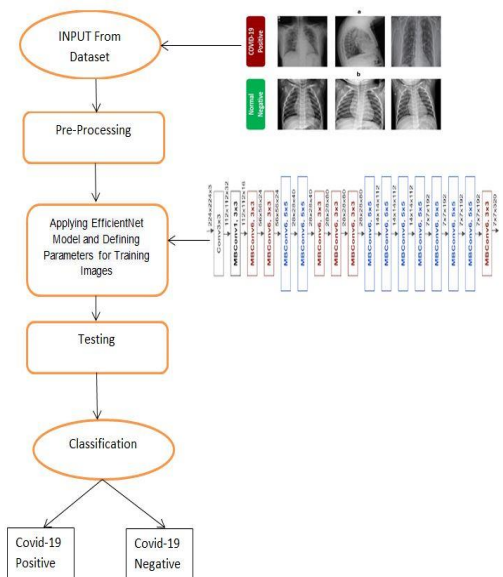


Fig. 4. Flow Chart Representation of Proposed Model.



#### IV. EVALUATION AND RESULT

There are various methods we can apply for the Loss calculation. Here in the proposed method cross-entropy loss is used for the calculation of the losses because of the chosen classification. In the similar manner for the weight update Adam optimizer has chosen which is well suited for deep learning applications. Proposed network trained with above mentioned parameters and the number of epochs is set to 10, out of which the network attained an accuracy of 99%. Training model attained the accuracy for both the classes as 99%-100%.

Fig. 5 shows the Graph between Accuracy and Loss. There is over 7000+ batch iterations taken place and the plot showcases these changes occurring in the values with respect to the batch iterations.

After the training is performed now the trained model has to be applied for the testing samples. Now the transforms are composed and applied to the test dataset to make it conform to the same distribution as the training dataset. Proposed model performed prediction of the test image with an accuracy of 99%. With the proposed model, a medical practitioner can identify Covid accurately in 99 out of 100 people rapidly.

Comparison of the proposed model results with the few of the existing models are shown in Table I.

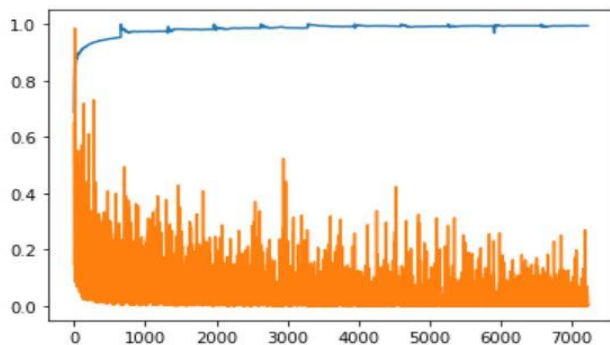


Fig. 5. Plot between Accuracy and Loss.

TABLE I. COMPARISON OF DIFFERENT MODEL ACCURACIES

Authors	Imaging Type	Dataset Size (No. of Images)	Model Used	Accuracy
Boran Sekeroglu et al.	CXR	6100	CNN	98.5%
Thiyagarajan Padmapriya et al.	CXR & CT	650 349	Multimodal covid network	99.75%
Yu-Huan Wu et al.	CT	3855	JCS	95.9%
Guangyu Jia	CXR & CT	7592	Dynamic CNN	99.3%
Bhawna Nigam	CXR		EfficientNet	93.4%
Dandi Yang	CT	2481	VGG 16	99%
Proposed Method	CXR	15000+	Efficient b3	99%

#### V. CONCLUSION

The most important factor in life is health; therefore early diagnosis of any disease is required. In this current pandemic kind situations early and accurate diagnosis of Covid-19 from radiological examinations like CT scan and Radiograph images is very important for Doctors as well as patients. Proposed model implemented with Deep learning approach based on X-ray images to reduce the financial cost and radiation effect to the patients as well as to diagnose in less time. The proposed model is capable of classifying the radiograph images as Covid positive or negative in less time with an accuracy of 99%. The model trained on huge data samples and attained an accuracy of 100% in training and 99% in testing. This way the proposed approach constructed with CNN architecture on EfficientNet can provide desired assistance in diagnosing Covid-19 for radiologists.

#### REFERENCES

- [1] Coronavirus Disease 2019. 2020. [Online]. Available: [https://en.wikipedia.org/wiki/Coronavirus\\_disease\\_2019](https://en.wikipedia.org/wiki/Coronavirus_disease_2019).
- [2] Shen, M., Zhou, Y., Ye, J., Abdullah AL-maskri, A. A., Kang, Y., Zeng, S., & Cai, S. (2020). Recent advances and perspectives of nucleic acid detection for coronavirus. *Journal of Pharmaceutical Analysis*, 10(2), 97–101. <https://doi.org/10.1016/j.jpaha.2020.02.010>.
- [3] Tahamtan, A., & Ardebili, A. (2020). Real-time RT-PCR in COVID-19 detection: Issues affecting the results. *Expert Review of Molecular Diagnostics*, 20(5), 453–454. <https://doi.org/10.1080/14737159.2020.1757437>.
- [4] Wong, H. Y., Lam, H. Y., Fong, A. H.-T., Leung, S. T., Chin, T. W.-Y., Lo, C. S., Lui, M. M.-S., Lee, J. C., Chiu, K. W.-H., Chung, T. W.-H., Lee, E. Y., Wan, E. Y., Hung, I. F., Lam, T. P., Kuo, M. D., & Ng, M.-Y. (2020). Frequency and distribution of chest radiographic findings in patients positive for covid-19. *Radiology*, 296(2). <https://doi.org/10.1148/radiol.2020201160>.
- [5] Li, Y., Yao, L., Li, J., Chen, L., Song, Y., Cai, Z., & Yang, C. (2020). Stability issues of RT - PCR testing of SARS - COV - 2 for hospitalized patients clinically diagnosed with COVID - 19. *Journal of Medical Virology*, 92(7), 903-908. <https://doi.org/10.1002/jmv.25786>.
- [6] Borghesi, A., & Maroldi, R. (2020). Covid-19 outbreak in Italy: Experimental chest X-ray scoring system for quantifying and monitoring disease progression. *La Radiologia Medica*, 125(5), 509–513. <https://doi.org/10.1007/s11547-020-01200-3>.
- [7] A. Jacobi, M. Chung, A. Bernheim, and C. Eber, “Portable chest x-ray in coronavirus disease-19 (COVID-19): A pictorial review,” *Clin. Imag.*, vol. 64, pp. 35–42, 2020.
- [8] Fang, Y., Zhang, H., Xie, J., Lin, M., Ying, L., Pang, P., & Ji, W. (2020). Sensitivity of chest CT for covid-19: Comparison to RT-PCR. *Radiology*, 296(2). <https://doi.org/10.1148/radiol.2020200432>.
- [9] Lakhani, P., & Sundaram, B. (2017). Deep learning at chest radiography: Automated Classification of pulmonary tuberculosis by using convolutional neural networks. *Radiology*, 284(2), 574–582. <https://doi.org/10.1148/radiol.2017162326>.
- [10] F. Shi et al., “Review of artificial intelligence techniques in imaging data acquisition, segmentation and diagnosis for COVID-19,” *IEEE Rev. Biomed. Eng.*, vol. 14, pp. 4–15, 2021, doi: 10.1109/RBME.2020.2987975.
- [11] Wynants, L., Van Calster, B., Collins, G. S., Riley, R. D., Heinze, G., Schuit, E., Bonten, M. M., Dahly, D. L., Damen, J. A., Debray, T. P., de Jong, V. M., De Vos, M., Dhiman, P., Haller, M. C., Harhay, M. O., Henckaerts, L., Heus, P., Kammer, M., Kreuzberger, N., ... van Smeden, M. (2020). Prediction models for diagnosis and prognosis of covid-19: Systematic Review and Critical Appraisal. *BMJ*, m1328. <https://doi.org/10.1136/bmj.m1328>.
- [12] Chowdhury, M. E., Rahman, T., Khandakar, A., Mazhar, R., Kadir, M. A., Mahbub, Z. B., Islam, K. R., Khan, M. S., Iqbal, A., Emadi, N. A., Reaz, M. B., & Islam, M. T. (2020). Can ai help in screening viral and



- covid-19 pneumonia? IEEE Access, 8, 132665–132676. <https://doi.org/10.1109/access.2020.3010287>.
- [13] M. Farooq and A. Hafeez, “COVID-ResNet: A deep learning framework for screening of COVID19 from radiographs,” 2020, arXiv:2003.14395.
- [14] Abdel-Basset, M., Chang, V., Hawash, H., Chakraborty, R. K., & Ryan, M. (2021). FSS-2019-nCov: A deep learning architecture for semi-supervised few-shot segmentation of COVID-19 infection. Knowledge-Based Systems, 212, 106647. <https://doi.org/10.1016/j.knsys.2020.106647>.
- [15] Maghdid, H., Asaad, A. T., Ghafoor, K. Z., Sadiq, A. S., Mirjalili, S., & Khan, M. K. (2021). Diagnosing covid-19 pneumonia from X-ray and CT images using Deep Learning and transfer learning algorithms. Multimodal Image Exploitation and Learning 2021. <https://doi.org/10.1117/12.2588672>.
- [16] Wang, L., Lin, Z. Q., & Wong, A. (2020). Covid-net: A tailored deep convolutional neural network design for detection of COVID-19 cases from chest X-ray images. Scientific Reports, 10(1). <https://doi.org/10.1038/s41598-020-76550-z>.
- [17] Kanumuri, C., & Madhavi, C. H. R. (2022). A survey: Brain tumor detection using MRI image with deep learning techniques. Smart and Sustainable Approaches for Optimizing Performance of Wireless Networks, 125–138. <https://doi.org/10.1002/9781119682554.ch6>.
- [18] S. Rajaraman, J. Siegelman, P. O. Alderson, L. S. Folio, L. R. Folio, and S. K. Antani, “Iteratively pruned deep learning ensembles for COVID-19 detection in chest x-rays,” 2020, arXiv:2004.08379.
- [19] Narin, A., Kaya, C., & Pamuk, Z. (2021). Automatic detection of coronavirus disease (COVID-19) using X-ray images and deep convolutional neural networks. Pattern Analysis and Applications, 24(3), 1207–1220. <https://doi.org/10.1007/s10044-021-00984-y>.
- [20] M. Z. Alom, M. Rahman, M. S. Nasrin, T. M. Taha, and V. K. Asari, COVID\_MNet: Covid-19 detection with multi-task deep learning approaches, 2020, arXiv:2004.03747.
- [21] <https://ai.googleblog.com/2019/05/efficientnet-improving-accuracy-and.html>.
- [22] Kandel, I., & Castelli, M. (2020). The effect of batch size on the generalizability of the Convolutional Neural Networks on a histopathology dataset. ICT Express, 6(4), 312–315. <https://doi.org/10.1016/j.icte.2020.04.010>.
- [23] <https://towardsdatascience.com/complete-architectural-details-of-all-efficientnet-models-5fd5b736142>.

# Machine Learning based Electromigration-aware Scheduler for Multi-core Processors

Jagadeesh Kumar P, Mini M G  
Department of Electronics Engineering  
Government Model Engineering College  
Thrikkakara, Cochin 21  
India

**Abstract**—The rising performance demands in modern technology devices see the need to pack more functionality per area and are made possible with the advent of technology scaling. The extremely down-scaled, high-density processors used in such technology devices functioning at high frequencies and greater temperatures expedite various aging effects which impact the reliable lifetime of computing systems. Electromigration is considered to be an important intrinsic aging effect that reduces the useful lifetime of modern microprocessors. The objective of this work is to use machine learning methods to develop an electromigration-aware scheduler for assigning workloads to cores based on reliability and performance requirements. Aging estimation of the processor cores is performed based on the proposed computationally efficient and accurate regression-based thermal prediction models. According to experimental findings, the suggested technique can significantly extend the lifetime of multi-core architectures while allowing performance to degrade gracefully. The maximum error in the prediction of the lifetime of the cores using the proposed methodology is estimated to be 2.85%.

**Keywords**—*Electromigration aware scheduler; useful lifetime; multi-core processor reliability; machine learning model*

## I. INTRODUCTION

The computing requirements of modern embedded systems in application areas including automotive, storage, networking, and 5G, among others, necessitate the use of high-performance processors. To deliver higher functionality per area demand for such applications, manufacturers create dense integrated multi-core processor chips operating at higher speeds [1]. Increased power density and operating temperature of processor cores can expedite aging phenomena including electromigration, lowering the processor's quality and reliability [2]. Runtime task execution management is a pressing research topic to minimize such aging effects in processor cores [2], [3], [4]. A rise of 10 - 15°C in the operating temperature could reduce the processor lifespan to half [5]. The lifetime of multi-core processors can be improved by estimating the aging effects of tasks that are ready for execution. The workload assignment to cores and frequency adjustments can be done at runtime to minimize the effect of aging of the cores. Such strategies need to be computationally precise and quick for real-time implementation.

Transistor aging is considered an important phenomenon that affects the reliability and lifetime of modern CMOS integrated circuits [6]. Time-Dependent Dielectric Breakdown

(TDDB), Hot Carrier Injection (HCI), and Negative Bias Temperature Instability (NBTI) are the key aging effects that challenge the reliable operation of modern integrated circuits. The intrinsic effects, Electromigration (EM) and Stress migration (SM), which are due to the interconnect aging, are also significant in present CMOS technology devices [7]. New methods for investigating and predicting degradation effects are important for enhancing the lifetime reliability of modern CMOS technology devices.

The suggested method to increase the multi-core processors' lifespan reliability is presented in two sections. In the first section, machine learning models based on regression to predict the processor cores' steady state temperature for the incoming jobs are proposed. A fine-grained approach is followed in this work for predicting the thermal properties of the cores' processing components. The fine-grained approach is more suitable for understanding the localized characteristics of the cores. In the second section, a scheme for mitigating the aging effects of processor cores by implementing a runtime frequency control policy is presented, which takes the temperature estimates of the proposed machine learning models as input. The operating speeds of the cores are decided to take into account the aging effects and the performance requirements. Our experimental results show that the proposed aging-aware scheduler, which is guided by the developed machine learning models could predict the core's lifetime with a maximum error of 2.85 %. The average error in temperature estimation is assessed, for each of the proposed models, and the maximum value is observed as 0.492 °C.

The rest of the paper's contents are arranged in the order mentioned below. A number of the significant efforts on thermal prediction and the regulation of thermal challenges of the processor cores are discussed in Section II. Linear regression and polynomial regression schemes for the estimation of the core temperature and the concept of the proposed aging-aware scheduler for enhancing processor core lifetime reliability are presented in Section III. An analysis of the results obtained for the proposed schemes are mentioned in Section IV. Finally, Section V concludes the work and mentions the possibilities for further research in the field.

## II. LITERATURE REVIEW

The increasing performance demands of current technology gadgets have prompted various studies to concentrate on optimal job scheduling methods in multi-core systems

operating under real-time constraints. A significant body of published works focuses on techniques for a better performance-power trade-off of multi-core processors [8], [9], [10]. The effect of using different styles of coding for balancing between performance and energy consumption of the processing cores is presented in [11]. A core allocation technique to lower the energy usage of mobile devices by engaging the LITTLE cores to a maximum extent and ensuring the performance of the device is proposed in [12]. The execution time of complex programs can be minimized and thus the performance of the applications can be improved by the use of today's high-performance computing systems employing multi-core processors. Parallelization schemes [13] can be employed for converting the serial execution of programs into a hybrid parallel mode to take the advantage of the processing capability of multi-core processors. Reliability-aware scheduling techniques [14] can be used to reduce soft errors of heterogeneous chip-multiprocessors. The performance and power efficiency in heterogeneous multi-core CPUs can be enhanced with smart workload schedulers [15].

An aging-aware scheduler needs an accurate thermal estimate of the various logical components to make run-time decisions when a workload gets executed in a processor core. Several research works have looked into the feasibility of using model-based methodologies to determine the CPU core temperature characteristics [16], [17], [18]. An architectural level thermal behavioral modeling technique known as Thermsid [19] builds temperature models from the observed temperature and power statistics. Multiple scheduling schemes based on the efficient and simple thermal model are used [20] for managing the operations of homogeneous processor platforms. Thermal Estimation Accelerator (TEA) [21], a processing element level monitoring scheme for the temperature at runtime using hardware accelerators can serve as a benchmark for Dynamic Thermal Management (DTM) methods.

For the development of thermal models, it is required to estimate thermal profiles of the benchmark tasks executing on selected cores at defined operating frequencies. HotSpot [22] is a popular tool used in the temperature estimation of processor cores and is helpful in architectural studies. Various logical components in the processor architecture are represented as its equivalent thermal resistor and capacitor values along with the thermal package information [23]. Information about the floor design and power estimations for the logical components are supplied to HotSpot to determine the temperature profile. The power estimation of the logical components can be done with the tool McPAT (Multi-core Power, Area, and Timing) [24]. Based on high-level data, such as the frequency of operation of the core, McPAT can predict the architectural level power usage of the processor core containing caches and memory controllers. A McPAT-monolithic framework is presented in [25] for the architecture modeling of 3-D hybrid monolithic multi-core systems. The work in [26] proposes a micro-architectural framework to estimate the performance and energy consumption of cores in a multi-core processor. A detailed validation of McPAT's power models done with the help of a toolchain used in industrial practice is presented in [27]. In this study, McPAT is utilized to calculate the dynamic

power of the core's logical subsystems. McPAT requires the operation statistics of the applications and micro-architectural characteristics as its inputs to calculate the power consumption of every system component. The operation statistics of the tasks can be evaluated using the gem5 simulator [28]. CPU models with different types of memory configurations and cache organizations can be defined for the analysis. The current generation of widely used commercial Instruction Set Architectures (ISAs) including ARM and x86 are supported by gem5. A significant number of these simulators are utilized in the relevant fields of research since they are useful in analyzing the performance and power consumption of various processor models and can be used to validate various design options. A study of the basics of several multiprocessor simulation methodologies and a summary of the correctness of six architecture simulators including gem5 are presented in [29]. Gem5-X, a framework for system-level simulation based on gem5 [30] may be employed to assess the potential benefits of the architectural extensions for many image processing-related applications.

Computer architecture simulators can be used to closely examine the execution properties of applications that run in a processing core. Regression-based models with high computational efficiency and accuracy may well be constructed by relating the thermal figures estimated with HotSpot to the application characteristics estimated using gem5. Representative applications in open-source benchmark MiBench [31] can be used to develop and validate thermal models of typical workloads running in embedded processors. MiBench presents a collection of 35 embedded programs organized in six categories, each of which focuses on a specific segment of the embedded market. Workloads belonging to the application areas: network, security, telecommunication, and consumer from MiBench suite are used in this work. The characteristics of the tasks jpeg encode/decode, 32-bit Cyclic Redundancy Check (CRC), Dijkstra, and Secure Hash Algorithm (SHA), representing the mentioned application areas, running on x86 architecture-based processing cores are analyzed using gem5 and are used to train the regression models. The MiBench suite's consumer device benchmarks are designed to simulate the consumer device applications found in products including Personal Digital Assistants (PDAs), scanners, and digital cameras. This category largely focuses on image processing, and one of the representative image compression and decompression technologies is JPEG encoding/decoding. The telecommunications category of applications stands close to consumer applications because of the increased demand for consumer devices with wireless communication capability. Cyclic redundancy checks (CRC) are often used in data transmission for the detection of errors. A 32-bit cyclic redundancy check is performed on a sound file from the adaptive differential pulse code modulation benchmark as part of the CRC32 test.

Devices such as switches and routers have embedded applications that fall under the network category. Finding the shortest path through a graph is one of the methods used to illustrate the networking category. The Dijkstra benchmark determines the shortest route across each set of nodes of a graph by repeatedly applying Dijkstra's algorithm. In

applications related to e-commerce, data security is a key factor. Security applications frequently employ a variety of hashing, encryption, and decryption algorithms. The Secure Hash Algorithm (SHA) is frequently used to create digital signatures and transfer cryptographic keys in a secure manner. The secure hashing method, SHA benchmark in MiBench, generates a 160-bit code for an input.

Estimating the temperature and power at runtime using the tools HotSpot and McPAT is computationally expensive and limits their implementation in real-time schedulers. A regression-based model can be developed based on the thermal and power profiles of the workloads estimated using HotSpot and McPAT. Such trained and created regression-based models can quickly and accurately estimate the temperature of the logical components, and they are suitable for the successful application of aging-aware scheduling methods. A method for constructing a compact machine learning-based thermal prediction system appropriate for fast decision-making is presented in [32]. A temperature control strategy based on machine learning to determine the appropriate core frequency and encoder configuration for High-Efficiency Video Coding (HEVC) is proposed in [33]. A machine learning and simulation-based approach can be employed [34] to estimate the temperature map of a chip using the power consumption, utilization of the core, and recorded sensor temperatures. As discussed, a significant amount of the related research works published in recent years propose different approaches for determining the critical parameters associated with the workloads using software tools. Many of these works emphasize the increasing need for runtime techniques to regulate the processor temperature by changing its operating behavior. Such recent research works propose many techniques for better performance-power trade-offs. But, in our understanding, the use of scheduling techniques to reduce aging with the help of computationally efficient runtime temperature estimation methods for high-performance applications running on multi-core processors is a largely unexplored topic.

### III. PROPOSED WORK

The first part of the proposed work deals with thermal profile modeling problems. Here, a regression-based model is employed to predict the steady-state temperature of the processing elements in a multi-core architecture. The temperature estimation model is driven by the properties of the workload. In the second part, an aging-aware scheduler is presented, whose scheduling activities are based on the thermal estimates of the regression model proposed in the first part of the work. The scheduler utilizes the thermal estimations and the performance need of the workloads taking into account the operating speed of the cores and trying to minimize the effect of processor aging.

#### A. Development of Regression-based Models

Regression analysis is one of the powerful multivariate statistical techniques to infer and form a functional relationship in a population [35]. In this work, regression analysis is used to relate the workload characteristics with the thermal effects of the processing elements. Workloads in embedded applications are suitable for implementing the suggested scheme because

they typically have high levels of predictability in their execution characteristics, such as the instruction execution behavior, memory operations, and the kind of information processing. The execution patterns of the MiBench applications running on a Hexa-deca homogenous multi-core architecture are analyzed using the gem5 simulator. The gem5 can be set up to operate in either Syscall Emulation (SE) mode or Full System (FS) mode. In SE mode, gem5 can imitate system calls made by applications. When configured in the FS mode, gem5 creates a bare-metal context for executing an operating system. In this work, gem5 is configured in the SE mode for analyzing the patterns of execution of the benchmark applications. The cores of the multi-core processor are selected as having x86 architecture. The cache hierarchy is defined to be of two levels, with level 1 private cache and level 2 shared cache. A subset  $j$  of the workload parameters is employed in this study, which directly affects the thermal profiles of the various logical components of the core. i.e.,  $j \in \{W\}$ , where  $W$  represents the complete set of workload characteristics analyzed using gem5.

The above-mentioned workloads from the MiBench suite, which cover several embedded application areas, are used for model development and analysis. The selected applications are analyzed using the gem5 simulation tool and the characteristics are determined. The power usage of the various functional parts of the CPU architecture is calculated using McPAT. The thermal model of HotSpot is driven by the estimated power traces, and the chip and package characteristics. The configuration file specifies the parameters of the processor core for the HotSpot tool, which are shown in Table I.

Linear Regression (LR) and Polynomial Regression (PR) are the two regression models used in this work for estimating the thermal values of the functional elements of the CPU. In the Linear Regression (LR) model, the steady-state temperature of a functional element is represented as a weighted sum of the selected workload characteristics. In the LR model, the predicted temperature is represented as in (1).

$$\hat{y}(w, x) = w_1 x_1 + \dots + w_p x_p + b \quad (1)$$

TABLE I. HOTSPOT CONFIGURATION PARAMETERS

HotSpot Parameters	Value
Thickness of the chip (in meters)	0.00015
Specific heat of Silicon (in J/(m <sup>3</sup> -K))	1.75 x 10e6
Thermal conductivity of Silicon (in W/(m-K))	100.0
Resistance (Convection) (in K/W)	0.1
Capacitance (Convection) in J/K (Heat sink)	140.4
Thickness (Heatsink) (in meters)	0.0069
Heatsink side (in meters)	0.06
Thermal conductivity of Heatsink (in W/(m-K))	400
Specific heat (Heatsink) (in J/(m <sup>3</sup> -K))	3.55 x 10e6
Side (Heat spreader) (in meters)	0.03
Thickness (Heat spreader) (in meters)	0.001
Thermal conductivity (Heat spreader) in W/(m-K)	400

where  $X = (x_1, x_2, \dots, x_p)$  represents the features used to train the models,  $w_1, w_2, \dots, w_p$  are the coefficients and  $b$  represents the bias. To reduce the sum of the squared estimate of errors between the measured values of the data, linear regression fits a linear model using weights  $W = (w_1, w_2, \dots, w_p)$ . The loss function which indicates the adequacy of the fit is given by (2).

$$L(\hat{y}, t) = \frac{1}{2}(\hat{y} - t)^2 \quad (2)$$

where  $\hat{y}$ ,  $t$ , and  $(\hat{y} - t)$  represent the prediction, target, and residual values respectively. The coefficients  $w_1, w_2, \dots, w_p$  and  $b$  are selected in a manner to reduce the loss function as represented in (3) and (4).

$$E(w_1, w_2, \dots, w_p, b) = \frac{1}{N} \sum_{i=1}^N L(y^i, t^i) \quad (3)$$

$$= \frac{1}{2N} \sum_{i=1}^N (\sum_j w_j x_j^i + b - t^i)^2 \quad (4)$$

In this study, the Python-based Scikit-learn machine learning package [36] is used, which is regarded as an effective and reliable tool for predictive data analysis.

In the development of the Polynomial Regression (PR) models of steady state temperature, polynomial regression needs to be performed on the data set, which are the workload characteristics, to fit a polynomial equation to it. This work extends linear regression by building polynomial features from the coefficients. For instance, the features in the second-order polynomials are utilized to fit a paraboloid to the data rather than a plane, giving the model represented in (5):

$$\hat{y}(w, x) = w_0 + w_1 x_1 + w_2 x_2 + w_3 x_1 x_2 + w_4 x_1^2 + w_5 x_2^2 \quad (5)$$

The above model can be considered as a linear model creating a set of features given in (6).

$$z = [x_1, x_2, x_1 x_2, x_1^2, x_2^2] \quad (6)$$

This renaming of the data allows for the formulation of the problem as in (7).

$$\hat{y}(w, x) = w_0 + w_1 z_1 + w_2 z_2 + w_3 z_3 + w_4 z_4 + w_5 z_5 \quad (7)$$

The derived polynomial regression belongs to a similar category of linear models as those previously mentioned, which can be evaluated using the same methods. The model is flexible enough to accommodate a broader range of data by taking into account the linear fits in a higher-dimensional space constructed with these basis functions. To transform an input data matrix into a new data matrix of a specific degree, the polynomial properties converter in the Scikit-learn Python machine learning toolkit is used. The parameters of  $X$  have been converted from  $[x_1, x_2, \dots]$  to  $[x_1, x_2, x_1 x_2, x_1^2, x_2^2, \dots]$  and are now applicable to any linear model.

Linear and polynomial regression are used to model the temperature profiles of the processing units of the core. Fig. 1 represents the architecture of the processor core considered in this work. The models thus developed are used to predict the steady-state temperature of the various processing elements in the architecture. The regression models presented in this work are most appropriate for multi-core systems executing embedded tasks since their task characteristics are often highly predictable. The proposed model-based prediction logic is

computationally efficient and is more suitable for real-time thermal estimation. The estimated data can be used by an aging-aware scheduler to determine the best course of action for controlling the temperature below threshold levels while maintaining performance goals and extending the useful lifetime of processor cores.

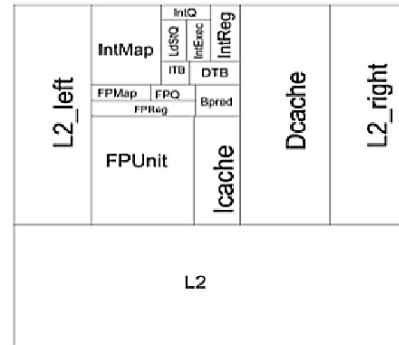


Fig. 1. Core Architecture.

### B. The Aging-aware Scheduler Design

With the advancements in integrated circuit design and fabrication technology, more transistors can now fit into a square millimeter of a silicon wafer. When clocked at higher speeds to address the execution constraints, such densely integrated processors have increased power and heat dissipation, which has a negative impact on lifetime dependability. The major challenge is to develop algorithms that can forecast device-level degradation behavior based on the application features. This work presents a strategy for extending the useful lifetime of multi-core processors. A fine-grained approach is followed for estimating the aging effects of the various processing components of the multi-core processor. The data available for the recent industrial-grade embedded processors manufactured by Texas Instruments [37] is taken as the reference. Referring to [37], the operational lifespan of the semiconductor core is taken as ten years, when the junction temperature  $T_j$  is  $105^\circ\text{C}$ . The crucial factor affecting silicon lifespan is the junction temperature  $T_j$  when the circuit is functioning within the limits of voltage and frequency stated in the data sheet.

Due to continual operation at high temperatures, wear-out processes begin to develop in semiconductor products during their useful life. The wear-out processes commonly considered in the design of integrated circuits include Gate Oxide Integrity (GOI) [38], Electromigration (EM) [39], [40], and Time-Dependent Di-electric Breakdown (TDDB) [41]. Additionally, the lifespan of the present semiconductor devices is affected by processes such as Negative Bias Temperature Instability (NBTI) [42] and Channel Hot Carriers (CHC) [43]. Among these, electromigration is a major aging effect in present integrated circuits. The primary factor that influences electromigration is the junction temperature  $T_j$ . The junction temperature  $T_j$  is thus the critical factor affecting silicon lifespan under electrical bias when the chip is operating within the prescribed data sheet conditions, and the lifetime can be represented using an Acceleration Factor (AF). The Arrhenius equation, which links the chemical reaction rate to temperature,

can be used to analyze the damage that occurs in electronic devices over time for various working temperatures. The accelerating factor (AF) [37] can be represented as in (8).

$$AF = \exp\left(\frac{E_a}{K} \left(\frac{1}{T_{use}} - \frac{1}{T_{stress}}\right)\right) \quad (8)$$

where AF represents the Acceleration Factor,  $E_a$  is the Activation energy in eV, K is the Boltzmann's constant ( $8.63 \times 10^{-5}$  eV/K),  $T_{use}$  is the use temperature in Kelvin and  $T_{stress}$  is the stress temperature in Kelvin.

This work proposes a methodology for improving the useful lifetime of the processor cores by considering electro-migration as the primary failure mechanism. The aging-aware scheduler estimates the temperature of the processing elements of the core when a job is ready, by using the workload characteristics as input to the developed models. Our goal is to allocate the tasks to cores based on the aforementioned inputs to maximize chip lifetime while meeting the performance requirement bounds. For the reliability-aware scheduler design, two types of frequency adaptations are proposed in this work, where the processor clock can either be controlled in discrete values or in a continuous manner. Based on the workload characteristics, the performance requirement, and the end system reliability requirement, the aging-aware scheduler selects a core from the pool of feasible cores and decides its frequency of operation. The architecture of the proposed aging-aware scheduler is represented in Fig. 2.

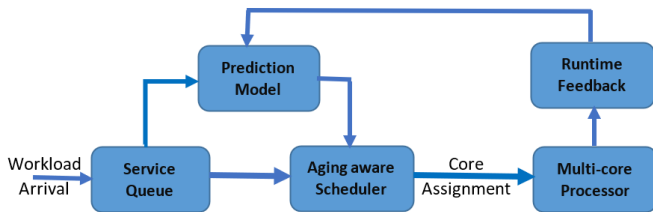


Fig. 2. The Aging-aware Scheduler.

The proposed aging-aware scheduler supports two types of clocking schemes for the processor core, i) cores whose operating frequencies can be selected from a discrete set of values and ii) cores whose frequency of operation can be varied continuously within a defined range. For the discrete frequency selection scheme, the scheduler will select the maximum possible operating frequency from a set of possible values, such that the lifetime reliability requirement can be met. In the continuous frequency selection scheme, the scheduler will have fine control of the operating frequency. In this case, the scheduler will use interpolate functions to estimate the closest frequency of operation that meets the lifetime reliability requirement with graceful performance degradation. For the specified workload characteristics, the data set includes known frequency and related temperature values of the logical components of the core.

Linear Interpolation (LI) and Spline Interpolation (SI) are the two types of interpolation schemes attempted in this work. To build a function utilizing fixed frequency datasets for linear interpolation, the `interp1d` class from the "scipy.interpolate" package is utilized. SciPy [44] is a free and open-source

Python library used for scientific and technical computing. When using linear interpolation for curve fitting, additional data sets are generated inside the boundaries of a finite collection of existing datasets using linear polynomials. In applications where smoothing is necessary, the piecewise-defined spline function is employed instead of polynomial interpolation because it produces good results for low-degree polynomials while minimizing Runge's phenomena for higher degrees.

Algorithm I illustrate the aging-aware scheduler's pseudo code. The scheduler, based on the workload characteristics, forms a feasible set of cores  $\{C_1, \dots, C_m\}$ , from the set of available cores during the scheduling interval. The temperature patterns of the logical components are predicted using the prediction models developed in the first part of this work, and using these parameters, the aging factor AF of each core  $C_i$  related to the lifetime reliability is determined. Based on the workload's performance requirements, the lifetime reliability requirement of the cores, and the type of frequency control supported by the architecture, i.e., either a discrete frequency control or fine frequency control, the frequency of operation of the core is determined.

---

#### Algorithm I. Aging Aware Core Selection

---

Inputs: workload characteristics, performance constraints, lifetime reliability requirement

1. **while** (true) {
2.   **for** each schedule window  $T_s$ , perform {
3.     **for** each task  $\{T_1, T_2, \dots, T_n\}$  in the process queue Q perform {
4.       a. analyze the characteristics of  $T_i$  and form the feasible set of cores  $\{C_1, \dots, C_m\}$ .
5.       b. estimate temperature characteristics of the processing elements  $\{L_1, \dots, L_p\}$  of the feasible cores  $C_i$ .
6.       c. estimate aging factor AF of each core  $C_i$  related to the lifetime reliability.
7.       d. select a core based on the lifetime reliability and performance requirements.
8.       e. determine the core's operating frequency  $f$ :
  - i. if discrete frequency control  
 $f = f_c$  where  $f_c \in \{f_1, \dots, f_s\}$
  - ii. else  $f = f_i$  where  $f_{min} \leq f_i \leq f_{max}$
9.     } //end for each task
10.   } //end for each schedule window
11. } //end while

Outputs: i) Mapping of the tasks  $\{T_1, \dots, T_n\}$  to cores  $\{C_1, \dots, C_m\}$  if  $n \leq m$ ; stall the remaining  $n-m$  tasks if  $n > m$ , ii) frequency of operation of the selected cores.

---

The regression models for predictive modeling need to be updated if the error in prediction is more than a threshold value because of the change in the data. The accuracy of the prediction logic is verified periodically with an updating interval. The updating interval, a customizable parameter, is kept substantially greater than the scheduling interval to reduce the computational overhead for the assessment of the actual temperature levels. The model updating process is represented in Algorithm II.



Algorithm II. Model Updation

Inputs: workload characteristics, core id, maximum allowable error in prediction (threshold)

1. **while** (true) {
2. **for** each updating interval,  $T_u$  do {
  - a. **for** each core  $\{C_1, \dots, C_m\}$  do {
    - compute the actual temperature of the processing elements (with on-chip sensors and/or software tools).
    - }//end** for each core
    - if** (prediction error > threshold)
      - update the prediction models.
    - end if**
  3. **}//end** for each updating interval ...
  4. **}//end** while

Output: updated prediction models.

IV. RESULTS AND DISCUSSION

The experiments of this research work were conducted using the application benchmarks belonging to consumer, telecommunications, network, and security categories taken from the well-known MiBench suite as represented in Table II.

Multiple instances created by altering the data set handled by the tasks are used to model the temperature characteristics of the logical components. As a result,  $n$  versions of a task  $w$  are created and evaluated its execution on  $m$  number of cores of the multi-core processor. Using new instances of the workloads derived from MiBench, the per-logical unit temperature is estimated to evaluate the developed models. The integer ALU, integer register file, floating-point unit, floating point register file, Data Translation Lookaside Buffer (DTLB), Instruction Translation Lookaside Buffer (ITLB), and load/store queue are characterized as the key power-consuming processing elements of the cores. Fig. 3 illustrates the validation of the developed models for the task CRC. The Steady State Temperature (SST) of the logical elements which are having significant power consumption is evaluated using the developed LR and PR models and is compared with the values estimated using the tool HotSpot. The operating frequency of the core is defined as 3400 MHz. The difference in estimation using the two methods is represented as a percentage error and is shown in Fig. 4. With a maximum prediction error of 0.008 percent for linear regression and 0.826 percent for polynomial regression-based models, the suggested regression-based models exhibit good consistency with HotSpot.

TABLE II. REPRESENTATIVE BENCHMARKS

Category	MiBench Benchmark
Consumer	JPEG encoding/decoding - (cjpeg /djpeg)
Telecom	Cyclic Redundancy Checks (CRC)
Network	Dijkstra
Security	Secure Hash Algorithm (SHA)

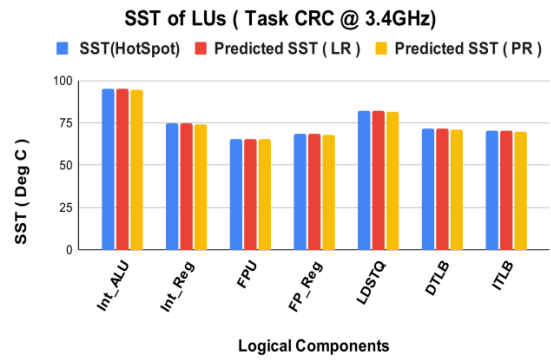


Fig. 3. SST of the Processing Elements Estimated using HotSpot, Linear Regression, and Polynomial Regression Models.

% error in the estimation of SST ( Task CRC @ 3.4GHz)

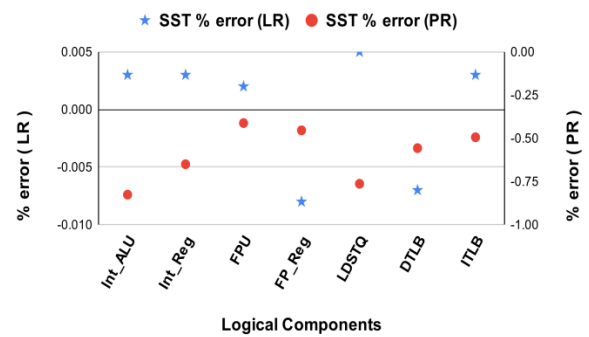


Fig. 4. Differences in the Estimation of SST of the Processing Elements are represented as a Percentage Error.

Fig. 5 illustrates the validation of the thermal models of the logical component integer ALU. The tasks used in the analysis are executed in cores set to operate at a clock frequency of 3400 MHz. The percentage error in the estimation of temperature is shown in Fig. 6. Simulation results show that the model is comparable to the HotSpot model in estimating the thermal profile of the logical components of the processor core.

SST of int\_ALU for different tasks ( core @ 3.4GHz)

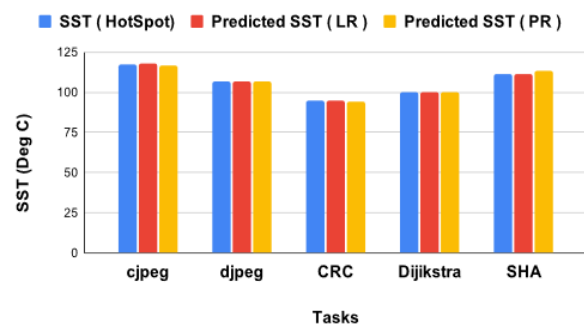


Fig. 5. Steady State Temperature of Integer ALU Estimated using HotSpot, Linear Regression, and Polynomial Regression Models.

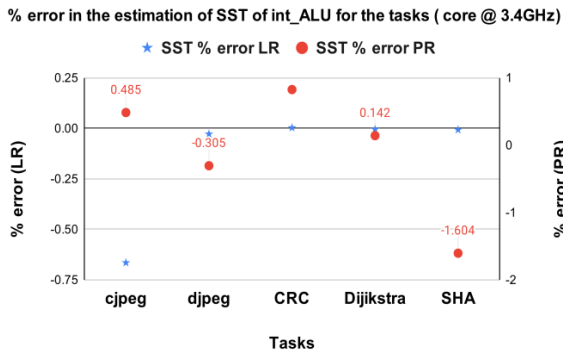


Fig. 6. Differences in the Estimation of Steady-state Temperature of Integer ALU.

The proposed aging-aware scheduler uses predicted SST of the logical components to estimate the degradation in the lifetime of the processor core during the scheduling of workloads. The lifetime of the critical components is represented using the Acceleration Factor (AF) while considering the junction temperature ( $T_j$ ) of silicon as the primary variable impacting the lifetime of the cores. Fig. 7 shows the validation of AFs of the principal power-consuming processing elements of the core for the task djpeg where the AFs are computed using the SST values estimated using HotSpot, LR, and PR models. Fig. 8 represents the validation of the AF estimation of int\_ALU for the different tasks. The processor core is set to operate at a frequency of 3400 MHz.

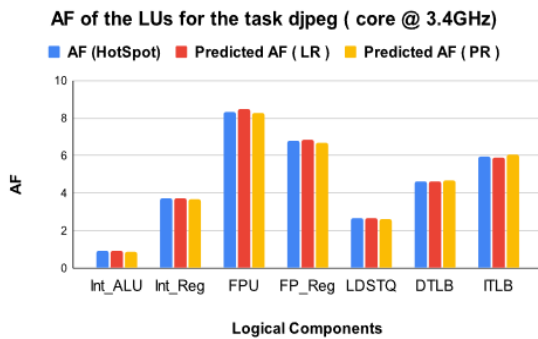


Fig. 7. Acceleration Factor of the Processing Elements Estimated using HotSpot, Linear Regression, and Polynomial Regression Models.

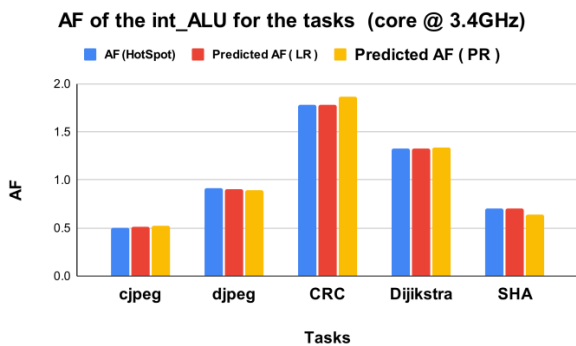


Fig. 8. Acceleration Factor of the Tasks Estimated using HotSpot, Linear Regression, and Polynomial Regression Models.

The proposed aging-aware scheduler estimates the temperature profiles of the logical components of the feasible cores based on the characteristics of the task in the service queue and decides the appropriate frequency of operation of the core. The lifetime reliability requirement of the core and the workload's performance requirements are used to determine the core operating frequency. In this work, the performance of the workloads running on a core clocked at 4000 MHz is taken as the reference, and the performance of the cores executing workloads with different operating conditions is represented relative to the reference performance. Fig. 9 represents the lifetime improvement of the cores and the corresponding relative performance degradation of the tasks when implementing a discrete frequency control scheme. In this case, the lifetime reliability requirement is taken as ten years, corresponding to an AF of 1. The core operating frequency is selected from the set {4GHz, 3.4GHz, 2.4GHz, 1.4GHz} based on the lifetime reliability requirement. Fig. 10 shows the corresponding values for AF = 1.5. The scheduler uses the thermal profile of the logical component having the highest value, for the estimation of AF of the core. This work assumes that the CPU core is designed to function for ten years when the junction temperature is at 105°C.

When the reliability-aware scheduler functions in the continuous frequency mode, the core operating frequency can assume a value within the defined range of 1400MHz to 4000 MHz. Linear Interpolation (LI) and Spline Interpolation (SI) are the two interpolation schemes employed, for estimating the closest frequency required for meeting the reliability and performance requirements. The frequency values for interpolation are determined based on the estimated temperature of the cores. Thermal estimation is performed using HotSpot, the standard method, and with the proposed LR and PR methods. The frequencies estimated by the scheduler for these temperature values are shown in Table III. The frequencies are determined for meeting the lifetime reliability requirement of ten years (corresponding AF = 1). Lifetime reliability of ten years corresponds to a threshold value of the silicon junction temperature,  $T_{qual}$  of 105 Degree Celsius.

Useful lifetime improvement and relative performance of tasks for AF = 1

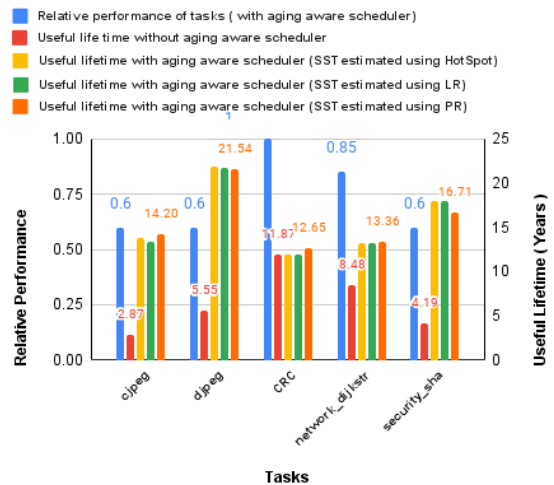


Fig. 9. Useful Lifetime Improvement and Relative Performance of Tasks (Discrete Frequency Control with AF = 1).

Useful lifetime improvement and relative performance of tasks for AF = 1.5

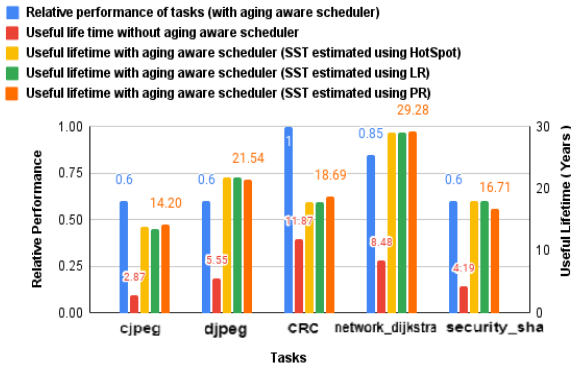


Fig. 10. Useful Lifetime Improvement and Relative Performance of Tasks (Discrete Frequency Control with AF = 1.5).

TABLE III. CORE OPERATING FREQUENCIES FOR AF=1

Fine Control of Operating Frequency (AF = 1 T <sub>qual</sub> = 105 Degree Celsius)						
Case	Case 1	Case 2	Case 3	Case 4	Case 5	Case 6
Temp Estimation scheme	Hot-Spot	Hot-Spot	LR	LR	PR	PR
Freq estimation scheme	LI	SI	LI	SI	LI	SI
Tasks	Estimated Frequencies of the cores (in GHz)					
cjpeg	2.7127	2.7128	2.7357	2.7356	2.7381	2.7377
djpeg	3.2871	3.2872	3.2852	3.2851	3.2663	3.2663
CRC	4.0	4.0	4.0	4.0	4.0	4.0
Dijkstra	3.7752	3.7752	3.7750	3.7749	3.7876	3.7876
SHA	3.0191	3.0195	3.0192	3.0191	2.9254	2.9253

Validation of the proposed scheme is carried out for the case where there is a lifetime reliability requirement of ten years. In this case, the aging-aware scheduler will adjust the operating frequency to limit the core temperature to 105°C. The actual temperature of the cores for the frequencies of operation mentioned in Table III is computed using HotSpot and is shown in Table IV. It can be seen that the proposed algorithm is adjusting the core frequencies in such a way that the temperature of operation of the cores is very close to the required value of 105°C. The average prediction error of the proposed scheme in the estimation of the core temperature is compared with the results reported in recent publications and is shown in Table V.

The lifetime of the cores, when operating with the temperatures shown in Table IV, is computed using (8). The theoretical lifetime corresponding to AF=1 is ten years. Table VI shows the lifetime of the cores corresponding to the steady state temperature values mentioned in Table IV. In this case, the algorithm is driving the operating frequency of the cores in such a way as to meet the lifetime requirement of ten years. The lifetime of the cores when operating with the frequencies estimated by the algorithm is having a maximum deviation of 2.85 % from the required lifetime.

TABLE IV. VALIDATION OF THE CORE TEMPERATURES FOR AF=1

Tasks	Temperature of Cores (in Degree Celsius) - Validation					
	Case 1	Case 2	Case 3	Case 4	Case 5	Case 6
cjpeg	104.99	104.99	105.4	105.4	105.46	105.44
djpeg	105.01	105.01	104.98	104.98	104.69	104.69
CRC	105.0	105.0	105.0	105.0	105.0	105.0
Dijkstra	105.0	105.0	104.98	104.98	105.15	105.15
SHA	105.0	105.0	105.0	105.0	103.46	103.46

TABLE V. COMPARISON OF THE PREDICTION ERROR OF CORE TEMPERATURE WITH PROPOSED METHODS IN THE LITERATURE

Scenarios	Average Prediction Error
Case 1	0.004 °C
Case 2	0.004 °C
Case 3	0.088 °C
Case 4	0.088 °C
Case 5	0.492 °C
Case 6	0.488 °C
Alzemiro et al. <sup>[21]</sup>	0.020 °C
Kaicheng Zhang et al. <sup>[32]</sup>	2.900 °C
Carlton Knox et al. <sup>[34]</sup>	1.390 °C

TABLE VI. THE ESTIMATED LIFETIME OF THE CORES

Tasks	Estimated Lifetime of Cores (in years) for AF=1					
	Case 1	Case 2	Case 3	Case 4	Case 5	Case 6
cjpeg	10.006	10.006	9.776	9.776	9.7426	9.7426
djpeg	9.9943	9.9943	10.011	10.011	10.178	10.178
CRC	10.0	10.0	10.0	10.0	10.0	10.0
Dijkstra	10.0	10.0	10.011	10.011	9.9152	9.9152
SHA	10.0	10.0	10.0	10.0	10.917	10.917

Using the methodology proposed, embedded application developers can perform a fast design space exploration between the lifetime reliability of the processor cores and the performance requirement of the tasks. Fig. 11 depicts the compromise between AF and the performance of the benchmark applications when linear regression is employed for temperature estimation along with linear interpolation for frequency estimation. At higher values of AF, processing cores have better lifetime reliability but at the expense of application execution performance. Fig. 12 represents the corresponding trade-off when polynomial regression is employed with continuous frequency assignment.

The concept suggested in this paper, where the aging-aware scheduler selects a task from a predetermined list of tasks to assign to a core, works well enough to extend the lifespan of multi-core systems running embedded workloads. A task might run at multiple scheduling points with a varying level of computational load because the complexity of the jobs that get executed on the cores may change over time. It is possible to account for these diverse computational costs at various execution times by building the regression model utilizing the heat profiles of logical units carrying out activities of varied computational costs at various execution times.

REFERENCES

- [1] V. Rajaraman, "Multi-core microprocessors," Resonance 22, no. 12, pp. 1175–1192, 2017, <https://doi.org/10.1007/s12045-017-0580-0>.
- [2] B. Wang, "Task Parallel Scheduling over Multi-core System," in M. G. Jaatun, G. Zhao, C. Rong, (eds), Cloud Computing, CloudCom 2009, Lecture Notes in Computer Science, vol 5931, Springer, Berlin, Heidelberg, [https://doi.org/10.1007/978-3-642-10665-1\\_39](https://doi.org/10.1007/978-3-642-10665-1_39).
- [3] S. Deniziak and A. Dzitkowski, "Scheduling of Distributed Algorithms for Low Power Embedded Systems," International Journal of Advanced Computer Science and Applications (IJACSA), 7(12), 2016. <http://dx.doi.org/10.14569/IJACSA.2016.071205>.
- [4] B. Ranjbar, T. D. A. Nguyen, A. Ejlali, and A. Kumar, "Power-Aware Run-Time Scheduler for Mixed-Criticality Systems on Multi-Core Platform," in IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, pp. 2009-2023, no. 10, 2020.
- [5] R. Viswanath, V. Wakharkar, A. Watwe, and V. Lebonheur, "Thermal Performance Challenges from Silicon to Systems", Intel Technology Journal, Q3 2000, pp. 1–16.
- [6] I. Hill, P. Chanawala, R. Singh, S. A. Sheikholeslam, and A. Ivanov, "CMOS Reliability From Past to Future: A Survey of Requirements, Trends, and Prediction Methods," in IEEE Transactions on Device and Materials Reliability, vol. 22, no. 1, pp. 1-18, March 2022, <https://doi.org/10.1109/TDMR.2021.3131345>.
- [7] M. D. Shroff and A. L. Loke, "Design-technology co-optimization for reliability and quality in advanced nodes," Proc. SPIE 11614, in Design-Process-Technology Co-optimization XV, 1161403, February 2021, <https://doi.org/10.1117/12.2585220>.
- [8] X. Yao, P. Geng, and X. Du, "A Task Scheduling Algorithm for Multi-core Processors," 2013 International Conference on Parallel and Distributed Computing, Applications and Technologies, pp. 259-264, 2013, <https://doi.org/10.1109/PDCAT.2013.47>.
- [9] T. Zhang, X. Pan, W. Shu, and M. Y. Wu, "Asymmetry-Aware Scheduling in Heterogeneous Multi-core Architectures," in C. H. Hsu, X. Li, X. Shi, R. Zheng, (eds), Network and Parallel Computing, NPC 2013, Lecture Notes in Computer Science, vol 8147, Springer, Berlin, Heidelberg, [https://doi.org/10.1007/978-3-642-40820-5\\_22](https://doi.org/10.1007/978-3-642-40820-5_22).
- [10] S. Holmbacka and J. Keller, "Workload type-aware scheduling on big.LITTLE platforms," in International Conference on Algorithms and Architectures for Parallel Processing, Springer, Cham, pp. 3-17, 2017, [https://doi.org/10.1007/978-3-319-65482-9\\_1](https://doi.org/10.1007/978-3-319-65482-9_1).
- [11] H. H. Hassan, A. S. Moussa, and I. Farag, "Performance vs. Power and Energy Consumption: Impact of Coding Style and Compiler," International Journal of Advanced Computer Science and Applications IJACSA, 8(12), 2017, <http://dx.doi.org/10.14569/IJACSA.2017.081217>.
- [12] D. Kim, Y. Ko, and S. Lim, "Energy-Efficient Real-Time Multi-Core Assignment Scheme for Asymmetric Multi-Core Mobile Devices," in IEEE Access, vol. 8, pp. 117324-117334, 2020, <https://doi.org/10.1109/ACCESS.2020.3005235>.
- [13] A. Algarni, A. Alofi, and F. Eassa, "Parallelization Technique using Hybrid Programming Model," International Journal of Advanced Computer Science and Applications (IJACSA), 12(2), 2021. <http://dx.doi.org/10.14569/IJACSA.2021.0120285>.
- [14] A. Naithani, S. Eyerman, and L. Eeckhout, "Reliability-Aware Scheduling on Heterogeneous Multicore Processors," 2017 IEEE International Symposium on High Performance Computer Architecture (HPCA), 2017, pp. 397-408, <https://doi.org/10.1109/HPCA.2017.12>.
- [15] S. I. Kim and J. Kim, "A Method to Construct Task Scheduling Algorithms for Heterogeneous Multi-Core Systems," in IEEE Access, vol. 7, pp. 142640-142651, 2019, <https://doi.org/10.1109/ACCESS.2019.2944238>.
- [16] Guoping Xu, "Thermal Modeling of Multi-Core Processors," Thermal and Thermomechanical Proceedings, 10th Intersociety Conference on Phenomena in Electronics Systems, 2006, IThERM 2006, pp. 96-100, <https://doi.org/10.1109/ITHERM.2006.1645327>.
- [17] D. Jaeckle and A. Sikora, "Thermal modeling of homogeneous embedded multi-core processors," 2014 International Conference on Advances in Computing, Communications and Informatics (ICACCI), pp. 588-593, 2014, <https://doi.org/10.1109/ICACCI.2014.6968448>.

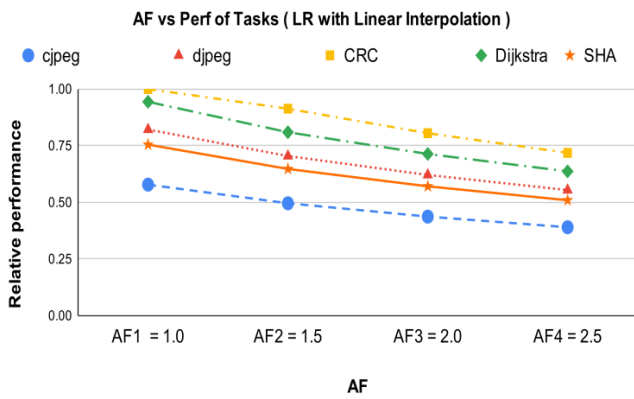


Fig. 11. Acceleration Factor (AF) - Performance Tradeoff (Linear Regression with Linear Interpolation).

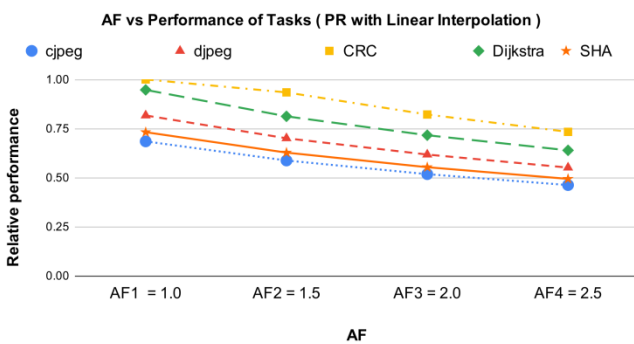


Fig. 12. Acceleration Factor (AF) - Performance Tradeoff (Polynomial Regression with Linear Interpolation).

V. CONCLUSION AND FUTURE WORK

The aging-aware scheduler proposed in this work uses the developed computationally efficient models to estimate the steady-state temperature of the processing elements in multi-core processor architecture. Temperature values estimated with the models are used to predict electromigration-induced aging. The scheduler performs an aging-aware application mapping strategy for enhancing the lifetime reliability of the cores. The suggested scheduler will estimate the operating frequency of the processing cores for satisfying the lifetime reliability constraints with a gentle decline of the performance as opposed to a no-aging aware scheduler, where the workloads are distributed to the cores based on the performance need. Results from simulations show that the suggested approach can increase the lifespan of the operation of multi-core processor systems.

The algorithm proposed in this work is extensible and configurable. The proposed framework is configurable, as it is possible to use on-chip thermal sensor data for estimating the temperature and aging effects of the logical components along with the temperature data computed using the software tools. In the future, the framework may be extended to take into account the impacts of aging brought on by Hot Carrier Injection (HCI), Positive-Bias Temperature Instability (PBTI), and Negative-Bias Temperature Instability (NBTI), along with electromigration.

- [18] I. Takouna, W. Dawoud, and C. Meinel, "Accurate Multicore Processor Power Models for Power-Aware Resource Management," 2011 IEEE Ninth International Conference on Dependable, Autonomic and Secure Computing, pp. 419-426, 2011, <https://doi.org/10.1109/DASC.2011.85>.
- [19] T. J. A. Eguia, R. Shen, S. X. D. Tan, E. H. Pacheco, and M. Tirumala, "Architecture level thermal modeling for multi-core systems using subspace system method," IEEE 8th International Conference on ASIC, pp. 714-717, 2009, <https://doi.org/10.1109/ASICON.2009.5351305>.
- [20] J. P. Rodríguez and P. M. Yomsi, "Work-in-Progress: Towards a fine-grain thermal model for uniform multi-core processors," 2020 IEEE Real-Time Systems Symposium (RTSS), pp. 403-406, 2020, <https://doi.org/10.1109/RTSS49844.2020.00049>.
- [21] A. L. da Silva, A. L. D. M. Martins, and F. G. Moraes, "Fine-grain temperature monitoring for many-core systems," in Proceedings of the 32nd Symposium on Integrated Circuits and Systems Design, pp.1-6, 2019, <https://doi.org/10.1145/3338852.3339841>.
- [22] R. Zhang, M. R. Stan, and K. Skadeon, "HotSpot6.0: Validation Acceleration and Extension," Tech. Report CS-2015-04.
- [23] M. R. Stan, K. Skadron, M. Barcella, W. Huang, K. Sankaranarayanan, et al., "HotSpot: a dynamic compact thermal model at the processor architecture level," Microelectronics Journal, Dec 2003, 34, (12), pp. 1153-1165, [https://doi.org/10.1016/S0026-2692\(03\)00206-4](https://doi.org/10.1016/S0026-2692(03)00206-4).
- [24] S. Li, J. H. Ahn, R. D. Strong, J. B. Brockman, D. M. Tullsen, et al., "McPAT: An integrated power, area, and timing modeling framework for multicore and manycore architectures," 42nd Annual IEEE/ACM International Symposium on Microarchitecture (MICRO), 12-16 Dec 2009, pp.469-480, <https://doi.org/10.1145/1669112.1669172>.
- [25] A. Guler and N. K. Jha, "McPAT-Monolithic: An Area/Power/Timing Architecture Modeling Framework for 3-D Hybrid Monolithic Multicore Systems," in IEEE Transactions on Very Large Scale Integration (VLSI) Systems, vol. 28, no. 10, pp. 2146-2156, Oct. 2020, <https://doi.org/10.1109/TVLSI.2020.3002723>.
- [26] F. A. Endo, D. Couroussé, and H. Charles, "Micro-architectural simulation of embedded core heterogeneity with gem5 and McPAT," Proceedings of the 2015 Workshop on Rapid Simulation and Performance Evaluation: Methods and Tools, RAPIDO '15, Amsterdam, Holland, Jan 2015, Article No.7.
- [27] S. L. Xi, H. Jacobson, P. Bose, G. Wei, and D. C. Brooks, "Quantifying sources of error in McPAT and potential impacts on architectural studies," IEEE 21st International Symposium on High Performance Computer Architecture (HPCA), CA, USA, Feb 2015, pp.577-589.
- [28] N. L. Binkert, B. M. Beckmann, G. Black, S. K. Reinhardt, A. G. Sadi, et al., "The gem5 Simulator," ACM SIGARCH Computer Architecture News, May 2011, 39, (2), pp. 1-7, <https://doi.org/10.1145/2024716.2024718>.
- [29] A. Akram and L. Sawalha, "A Survey of Computer Architecture Simulation Techniques and Tools," in IEEE Access, vol. 7, pp. 78120-78145, 2019, <https://doi.org/10.1109/ACCESS.2019.2917698>.
- [30] Y. M. Qureshi, W. A. Simon, M. Zapater, D. Atienza, and K. Olcoz, "Gem5-X: A Gem5-Based System Level Simulation Framework to Optimize Many-Core Platforms," 2019 Spring Simulation Conference, pp.1-12, 2019, <https://doi.org/10.23919/SpringSim.2019.8732862>.
- [31] M. R. Guthaus, J. S. Ringenberg, D. Ernst, T. M. Austin, T. Mudge, et al., "MiBench: A free, commercially representative embedded benchmark suite," Proceedings of the Fourth Annual IEEE International Workshop on Workload Characterization. WWC-4 (Cat. No.01EX538), 2001, pp. 3-14, <https://doi.org/10.1109/WWC.2001.990739>.
- [32] K. Zhang, A. Guliani, S. O. Memik, G. Memik, K. Yoshii, et al., "Machine Learning-Based Temperature Prediction for Runtime Thermal Management Across System Components," in IEEE Transactions on Parallel and Distributed Systems, vol. 29, no. 2, pp. 405-419, 2018, <https://doi.org/10.1109/TPDS.2017.2732951>.
- [33] A. Iranfar, M. Zapater, and D. Atienza, "Work-in-progress: a machine learning-based approach for power and thermal management of next-generation video coding on MPSoCs," 2017 International Conference on Hardware/Software Codesign and System Synthesis (CODES+ISSS), pp. 1-2, 2017, <https://doi.org/10.1145/3125502.3125533>.
- [34] C. Knox, Z. Yuan, and A. K. Coskun, "Machine Learning and Simulation Based Temperature Prediction on High-performance Processors," in Proceedings of ASME International Technical Conference and Exhibition on Packaging and Integration of Electronic and Photonic Microsystems (InterPACK), July 2022.
- [35] M. Mahbobi and T. K. Tiemann, "Introductory Business Statistics with Interactive Spreadsheets," 1st Canadian Edition, BCcampus, December 7, 2015, ch.8, Ebook ISBN 978-1-77420-007-0.
- [36] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, et al., "Scikit-learn: Machine Learning in Python," JMLR 12, pp. 2825-2830, 2011.
- [37] Allan Webber, "Calculating Useful Lifetimes of Embedded Processors," Texas Instruments Application Report, SPRABX4B, November 2014 Revised March 2020.
- [38] F. Bonoli, P. Godio, G. Borionetti, and R. Falster, "Gate oxide integrity dependence on substrate characteristics and SiO2 thickness," Materials Science in Semiconductor Processing, Volume 4, Issues 1-3, 2001, pp. 145-148, ISSN 1369-8001, [https://doi.org/10.1016/S1369-8001\(00\)00152-9](https://doi.org/10.1016/S1369-8001(00)00152-9).
- [39] J. R. Black, "Electromigration - A brief survey and some recent results," in IEEE Transactions on Electron Devices, vol. 16, no. 4, pp. 338-347, April 1969, <https://doi.org/10.1109/T-ED.1969.16754>.
- [40] D. G. Pierce and P. G. Brusius, "Electromigration: A review, Microelectronics Reliability," Volume 37, Issue 7, 1997, Pages 1053-1072,ISSN 0026-2714, [https://doi.org/10.1016/S0026-2714\(96\)00268-5](https://doi.org/10.1016/S0026-2714(96)00268-5).
- [41] J. F. Verweij and J. H. Klootwijk, "Dielectric breakdown: A review of oxide breakdown," Microelectronics Journal, Volume 27, Issue 7, 1996, pp. 611-622, ISSN 0026-2692, [https://doi.org/10.1016/0026-2692\(95\)00104-2](https://doi.org/10.1016/0026-2692(95)00104-2).
- [42] M. A. Alam and S. Mahapatra, "A comprehensive model of PMOS NBTI degradation," Microelectronics Reliability, 45:71-81, Aug. 2005.
- [43] Eiji Takeda, "Hot-carrier effects in scaled MOS devices," Microelectronics Reliability, Volume 33, Issues 11-12, 1993, Pages 1687-1711, ISSN 0026-2714, [https://doi.org/10.1016/0026-2714\(93\)90081-9](https://doi.org/10.1016/0026-2714(93)90081-9).
- [44] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, et al., "SciPy 1.0: fundamental algorithms for scientific computing in Python," Nat Methods 17, 261-272, 2020, <https://doi.org/10.1038/s41592-019-0686-2>.



# Sentiment Analysis on Acceptance of New Normal in COVID-19 Pandemic using Naïve Bayes Algorithm

Siti Hajar Aishah Samsudin<sup>1</sup>, Norlina Mohd Sabri<sup>2</sup>, Norulhidayah Isa<sup>3</sup>, Ummu Fatimah Mohd Bahrin<sup>4</sup>

Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA Cawangan Terengganu  
Kampus Kuala Terengganu, 21080 Kuala Terengganu, Malaysia

**Abstract**—The COVID-19 pandemic has such a significant impact and causes difficulties in many aspects that the new normal rules should be implemented to reduce the effects. New normal rules have been implemented by governments worldwide to break the virus chain and stop its transmission among the society. Even if the COVID-19 outbreak is under control, governments still need to know whether society could adapt and adjust to their new daily lifestyles. Many precautions still must be addressed as the transition to endemic status does not mean that COVID-19 will naturally eventually disappear. The World Health Organization also has warned that it is too early to treat COVID-19 as an endemic disease. Since the pandemic, many interactions have been done online, leading to the increasing social media usage to express opinions about COVID-19. The objective of the study is to explore the capability of the Naïve Bayes algorithm in the sentiment classification of the public's acceptance on the new normal in the COVID-19 pandemic. Naïve Bayes has been chosen for its good performance in solving various other classification problems. In this study, Twitter data were used for the analysis and were collected between March and June 2022. The evaluation results have shown that Naïve Bayes could generate excellent and acceptable performance in the classification with an accuracy of 83%. According to the findings of this research, many people have accepted the new normal in their daily lives. The future works would include scrapping more data based on geolocation, improving the feature extraction technique, balancing the dataset and comparing Naïve Bayes performance with other well-known classifiers. The subsequent study could also focus on detecting the emotions of the public and processing non-English tweets.

**Keywords**—Sentiment analysis; COVID-19; new normal; acceptance; naïve bayes

## I. INTRODUCTION

The term "sentiment" refers to a topic that includes subjective and objective aspects and factual and non-factual factors. It transcends the difference between a positive or negative subject [1]. Sentiment analysis is an analytical technique to analyze a text that identifies the level of public sentiment or opinion on a product or service and a person, such as politicians or celebrities [2]. The new normal is the new order, habits, and behavior based on adaptation to encourage clean and healthy living [3]. The Corona Virus Disease 2019 (COVID-19) pandemic has changed people's livelihood around the globe.

Consequently, COVID-19 causes so many difficulties to deal with in people's lives, that the governments need to implement new protocols to reduce the spread of the COVID-

19 infection among communities. The guidelines have become the new normal in the community's daily lives when people have suddenly been forced to adapt to all the protocols. Therefore, to mitigate the damaging effect of the COVID-19 pandemic, everyone is required to follow the Standard Operational Procedures set by the authorities in most countries. The new normal requires everyone to practice social distancing, use face masks, regularly wash hands with water and soap or sanitizer, stay at home unless necessary to go out, work from home or online learning for schools and universities [4]. Even after the infections have become less severe, people still have to take the precautions seriously.

It is necessary to break the virus chain and stop its transmission among the communities. Initially, the community was still ignorant of the virus's seriousness and was indirectly forced to adjust quickly and adapt to the new normal rules in daily life. Some people oppose the new normal life and still want to continue their old lifestyle without following health protocols and restrictions [5]. This kind of mentality may also influence others in embracing a new normal in their lives. Violations of health protocols will lead to an increase in COVID-19 cases. Implementing a new defence mechanism against a pandemic is quite challenging since it requires public engagement and acceptance of the policy required by the government [6]. In addition, due to the pandemic, any interaction was severely limited, resulting in increased digital use to obtain information about COVID-19. Therefore, social media sites such as Twitter have become essential platforms for expressing opinions, needs, and preferences. Nowadays, the community often responds to the current issues worldwide through Twitter by using tweets, retweeting others' posts, leaving a comment, or using a hashtag to spread something. A community has used Twitter to express their opinions over the increase of COVID-19 cases. The Twitter post can be a valuable source for understanding the community's acceptance towards new normal guidelines in making these new practices part of everyday habits.

As in Malaysia, after nearly two years of the pandemic, the country has entered the "Transition to Endemic" phase of COVID-10 starting from April 1, 2022, amid thousands of infections [7]. Endemic could be referred to as the constant presence, and usual prevalence of disease or infectious agent in a population within a geographic area. The transition to endemic can be considered an exit strategy.

Align with the announcement made by the Malaysia Government, this study proposes a sentiment analysis on the acceptance of the new normal in society. The tweet regarding



the transition phase will be used. The Naïve Bayes algorithm has been chosen as the machine learning method for sentiment classification. The Naïve Bayes algorithm would help to classify the level of sentiments of society's responses into two categories which are negative and positive. This study aims to explore the capability of the Naïve Bayes algorithm in the sentiment analysis on the acceptance of the new normal in the COVID-19 pandemic. The Naïve Bayes classifier has the advantage of requiring less training data to determine the estimated parameters needed in the classification process. Furthermore, the Naïve Bayes classifier is an algorithm frequently used for data mining because it is simple, fast to process, and easy to use with a simple model and a high-efficiency level [8].

This paper is structured as follows: Section I contains the Introduction; Section II discusses the Literature Review and Section III explains the Methodology. Section IV presents the Results and Discussion, while Section V presents the paper's Conclusion.

## II. LITERATURE REVIEW

### A. Similar Works

Several similar works are related to accepting the new normal in the COVID-19 pandemic. Table I describes the works, presenting the algorithms to solve the classification problems.

The first similar work used the Random Forest and Naïve Bayes algorithm to measure the people's sentiment toward government appeal in facing the COVID-19 pandemic [9]. Another work has implemented the Support Vector Machine Algorithm to analyze Twitter data and identify the Canadians' feelings regarding social distancing in relation to COVID-19 [10]. Research by [5] has analyzed the public's perception of social media towards the new normal during the COVID-19 pandemic in Indonesia. The study found that most Instagram users who follow religious accounts are against the new normal. The study [11] has implemented a Recursive Neural Network in analyzing the Twitter data to evaluate people's attitudes towards public health policies and events in the era of COVID-19. The tweet data analysis showed that many people's sentiments toward the stay-at-home approach were shifted because of the policy's negative consequences. Further, [12] has adopted the Latent Dirichlet Allocation (LDA) algorithm to perform sentiment, emotion, and content analysis of tweets regarding social distancing on Twitter. This research has indicated that most Twitter users supported the social distancing strategy.

In most works, the machine learning algorithms have solved the sentiment classification problems with reasonable accuracy. In this study, Naïve Bayes has been chosen due to its good performance in solving various other classification problems [13-15]. Although Naive Bayes has some drawbacks in the probability technique, it is worth exploring the algorithm's performance in solving another classification problem [16].

TABLE I. SIMILAR WORKS

	Title	Algorithm	Objective	Result	Ref
1.	Community Understanding of the Importance of Social Distancing Using Sentiment Analysis in Twitter	Random Forest Algorithm and Naïve Bayes algorithm	To measure people's sentiment toward government appeal in facing the COVID-19 pandemic.	Random Forest Algorithm had the best accuracy of 95.98%	[9]
2.	Sentiment Analysis on COVID-19-Related Social Distancing in Canada Using Twitter Data	Support Vector Machine algorithm	To analyse Twitter data and identify Canadians' feelings regarding social distancing.	SVM algorithm generated 87% of accuracy	[10]
3.	Public's Perception on social media towards New Normal during Covid-19 Pandemic in Indonesia: Content Analysis on Religious Social Media Accounts	Neuro-Linguistic Programming (NLP) method.	To discover about the public's perceptions of the government policies	The technique succeeds in solving the problem	[5]
4.	Analyzing Twitter Data to Evaluate People's Attitudes towards Public Health Policies and Events in the Era of COVID-19	Recursive Neural Network (RNN)	To track people's opinion regarding the public health policies and events during a COVID-19 pandemic.	The performance RNN is good.	[11]
5.	Understanding public perception of coronavirus disease 2019 (COVID-19) social distancing on Twitter	Latent Dirichlet Allocation (LDA)	To perform sentiment, emotion, and content analysis of tweets regarding social distancing on Twitter during the COVID-19 pandemic.	The algorithm able to generate good result	[12]

### B. Naïve Bayes Algorithm

The Naive Bayes classifier is a simple probabilistic classifier that calculates a set of probabilities by adding up the frequencies and value combinations from a given dataset. The algorithm uses the Bayes theorem and assumes that all variables are independently provided by the value of the class variable. The Naïve Bayes classifier can be trained very

effectively in supervised learning and can also be used in complicated real-life situations. The Naïve Bayes algorithm is simple to understand, requires training data to estimate the parameters, is unresponsive to unrelated features, and performs well when dealing with actual data and unique data source [17]. Below are the equations that calculate the probability categories in Naïve Bayes theorem.

$$P(c | x) = \frac{P(x | c)P(c)}{P(x)} \quad (1)$$

Equation (1) shows that variable c is class and variable x represents the attribute applied. P (c | x) is the posterior probability of class given attribute, P (x | c) is the likelihood, which is the probability of attribute given class, P (c) is a class prior probability. Lastly, P (x) is the predictor prior probabilities of the attribute [18].

### III. METHODOLOGY

#### A. Data Collection

The data collection were conducted from March to June 2022. The data were scrapped by using the Twitter API. A total of 7659 observations (rows) and three variables (columns) were obtained. This data collection process used English tweets comprising sentences or popular hashtags related to the new normal. The search keys were; “face mask”, “hand hygiene”, “Sejahtera scan”, “new normal COVID-19”, “stay at home”, “work from home” and “social distancing”. Table II shows the sample of a raw dataset from the scrapping process. The dataset contains three columns which represent Time, User and Tweet. However, only the Tweet column was used in this study.

#### B. Data Pre-processing

Data cleaning is the terms that refer to the process of identifying and correcting, removing, duplicate or invalid records from a database. Data inconsistency can occur in a variety of ways. For example, it might occur due to data corruption during transmission or storage or user entry errors [19]. Therefore, it is essential to clean up data so it can be used in models and produce better results.

TABLE II. SAMPLE OF RAW DATASET

Time	User	Tweet
2022-04-10 23:51:57+00:00	EnviroSmartGOP	#SocialDistancing , #lockdowns and changes to age-old ,PROVEN , #Quarantine methods of isolating sick , infectiousâ€ https://t.co/f7PShlaAOk
2022-04-10 23:22:20+00:00	MissyCooper13	RT @JohnSitarek: Mass PCR testing somewhere in #Eastworld. If you didn't have covid before standing in line for hours, you probably contracâ€
2022-04-13 17:20:01+00:00	NoxySA3	@theycrusjanssen you are reminded to get vaccinated, wear face mask and to use sanitizer to wash your hands #VaxxedForAfrica
2022-07-13 23:15:44+00:00	ThatTeddyH	RT @R_Chirgwin: Wear your masks. Avoid crowds. Work from home. Keep your children home. Drag these recalclitrant mendacious fools into line.â€

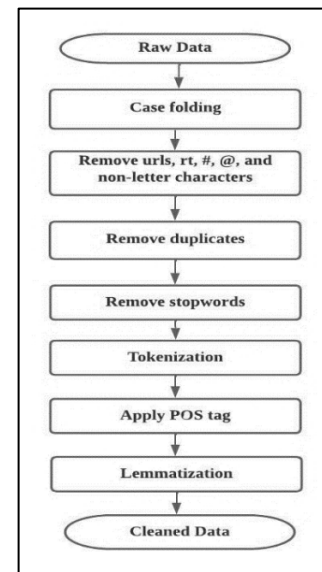


Fig. 1. Steps in Pre-processing.

Fig. 1 shows the steps of the data pre-processing for this study. Firstly, the tweets are converted into lowercase, while symbols and numbers are ignored. The website links such as “https” or “www”, retweets (rt), the hashtag symbol (#), user handles (@), and non-letter characters are also eliminated from the tweets. These are replaced with a blank string. After case folding, all the duplicate rows are removed to prevent redundancy in the dataset. Then, stopword removal is applied to the clean dataset, which removes stopwords listed in the NLTK package. To remove the stopwords, the lambda function is used. After stopwords removal, the tweet is tokenized. In tokenization, all text sentences are broken down into smaller parts called tokens. The Part-Of-Speech (POS) is applied after the data has been tokenized. The POS tagging determines the word class based on the word's placement in the sentences, indicating whether the word is a noun, adjective, verb, etc. and also enables future use of lemmatization.

The pos tags of a word are important to obtain the word's lemma properly. The final step in the process involves lemmatization steps being applied to the dataset. This is done because lemmatization has the potential to give meaningful root words. Lemmatization is preferred over stemming because it produces better results by performing an analysis based on the word's part and producing true dictionary words. After the preprocessing, the datasets were reduced to 2807 data. This is mainly because the pre-processing steps have eliminated all the noisy and unnecessary data.

#### C. Labelling

After the pre-processing, the processed tweets data must go through the labelling process. The labelling process is intended to label the data according to the sentiment classes, which is negative and positive [3]. Text Blob, a python library, has been used in this process. The positive tweets is represented by the number (+1), negative is represented by the number (-1) and neutral is represented by (0). In this study, only positive and negative classes are used. Table III shows an

example of tweets labelled with positive and negative sentiments. After the labelling, there are 2095 positive tweets, while the negative tweets obtained are 712.

TABLE III. EXAMPLE OF TWEETS LABELLED WITH POSITIVE AND NEGATIVE SENTIMENT

Tweet	Label
Keep hands clean, wear a mask for no more than a couple of hours and dont touch your face is the best advice.	1
I am sick of having to wear masks and have no face anymore. I don't want to wear a mask.	-1

D. Feature Extraction

Bag-of-Words is a method that has been used for feature extraction. Bag-of-Words is the most used technique for natural language processing. In this process, the bag-of-word extracts the words or the features from a tweet, and then the frequency of each term is calculated. It is called a “bag” of words because any information about the order or structure of words in the document is discarded. The model is only concerned with whether known words occur in the document, not where in the document [20].

As in this study, the bag of words model calculates the number of tokens collected in each document. Fig. 2 shows the output for the bag of words process. It displays the frequency table that counts how many times the term appeared. It contains three columns: the index representing the words, the count representing the occurrence of the terms, and the label representing each word's positive and negative labels.

E. System Architecture

The proposed system architecture is shown in Fig. 3. The first step is collecting the tweets from Twitter by using the Twitter API. The collected tweets are stored in the database and will go through the pre-processing steps. Then, the data is labelled with the positive or negative tag. Next, the data will be split into training and testing data. The training data will be used for feature extraction. The data is then passed to the Naive Bayes classifier model to categorize the data into positive or negative classes. The output result will show the accuracy of the proposed algorithm. Finally, the sentiment analysis result will be displayed to the user through the graphical user interface [21].

index	count	label
0	there	0 -1
1	nothing	6 -1
2	special	0 -1
3	born	0 -1
4	thing	16 -1
...	...	...
11639	gmfu	0 1
11640	realtor	1 1
11641	propey	1 1
11642	willing	1 1
11643	matt	1 1

Fig. 2. Bag of Words Model.

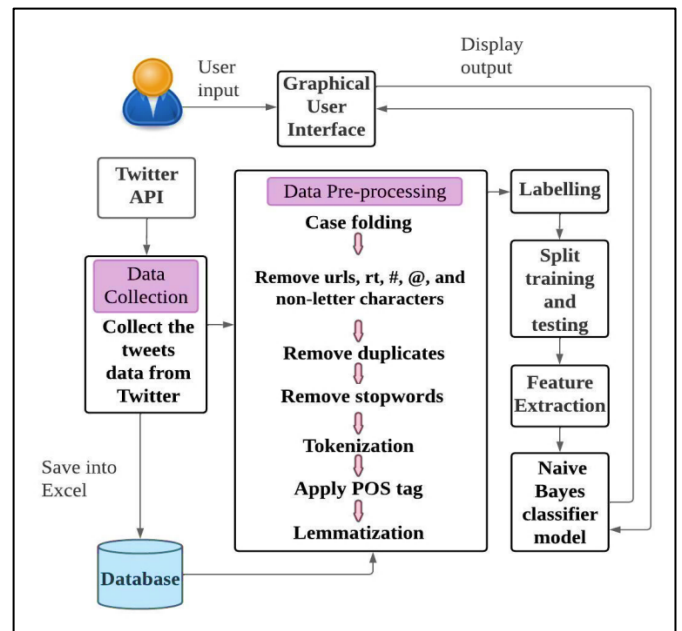


Fig. 3. Proposed System Architecture.

F. Performance Evaluation

Holdout method and Confusion Matrix have been used to evaluate the performance of the Naïve Bayes classifier. The holdout method is the simplest method to evaluate the performance of classifier where the data will be randomly split into two sets, which are training and testing set. The training data set is used to train the Naïve Bayes classifier and the testing dataset is used to test the performance of the classifier. Three sets were used in this study which are 90:10, 80:20 and 70:30. The first number for example for 90:10, indicates the percent of data used for training and the latter is for testing.

In addition, a confusion matrix is a matrix that comprises information on the actual and predicted classification achieved by classifier. It is often used to measure the performance of a classification algorithm. It includes the measurements of accuracy, precision, recall, F1-scores and ROC curve. The confusion matrix gives us a better picture of the algorithm's performance [22].

Table IV illustrates the confusion matrix for two classes which is for actual and predicted. The terms TP and TN indicate the True Positive and True Negative, which are referred to the accurately classified data. Meanwhile, FP and FN represent False Positive and False Negative, indicating incorrectly classified data [23].

TABLE IV. CONFUSION MATRIX

	Actual Positive	Actual Negative
Predicted Positive	TP	FP
Predicted Negative	FN	TN

The classifier accuracy can be calculated by applying the formula in (1). The accuracy presents the ratio of correct prediction over the total data. For the precision, it is indicated as the measure of the correctly identified positive cases from

all the predicted positive label. Next, for the recall, it measures the correctly classified positive from the total of the actual positive. As for the F1-score, it is the combination of precision and recall of a classifier into a single metric by using the harmonic mean [24]. The formulas to calculate the accuracy, precision, recall and F1-score are presented in the (1) to (5) respectively.

The accuracy obtained from (2) below:

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (2)$$

The Precision obtained from (3) below:

$$Precision = \frac{TP}{TP+FP} \quad (3)$$

The Recall obtained from (4) below:

$$Recall = \frac{TP}{TP+FN} \quad (4)$$

The F1-Score obtained from (5) below:

$$F1\ Score = 2 \times \frac{precision \times recall}{precision + recall} \quad (5)$$

Another classifier performance evaluation is the Receiver Operating Characteristic (ROC) curve. The ROC curve is a graph that summarizes the model's performance by integrating the confusion matrices at all threshold values [25]. Therefore, the ROC curve could provide an overview of the model's performance at different threshold values [26]. It is a graphical representation of the paired classifier with the bend indicating a trade-off between positive and false positive.

#### IV. RESULT AND DISCUSSION

There are two main analyses that have been conducted in this study. The first analysis conducted was the exploratory data analysis on the collected tweets data. Then, the second analysis was on the performance of the Naïve Bayes Classifier. In addition, the prototype to be used with the classifier model was proposed at the end of this section.

##### A. Exploratory Data Analysis

The first analysis is by analyzing the most common word obtained from the tweet. Fig. 4 shows the bar chart which is plotted to obtain the common words for the topics of acceptance of new normal in the COVID-19 pandemic. This chart provides an overview of which words frequently appear in the dataset. It reveals that the most common terms are new, home, normal, work, social, distancing, mask, wear, and stay. These top 10 common words have shown that people are aware of the pandemic's new normal lifestyle.

Then the analysis continues by analyzing the dataset according to its label which are positive and negative sentiment. The analysis conducted were word cloud and unigram analysis. A word cloud is one of the most common techniques for displaying and analyzing qualitative data. It is a graphic consisting of keywords found in the body of text, with the size of each keyword indicating the frequency with which it appears in the body of text [27].

Fig. 5 shows the word cloud for positive dataset, The most prominent words are "social", "distancing", "new", "normal",

"hand", "hygiene", "mask", "job", and "home". These words indicate the most discussed topic during the pandemic. It is used to represent positivity and actions during the pandemic. On the other hand, Fig. 6 the word cloud for negative dataset. We can observe some of the negative words such as "ill", "don't", "sick", "hate", "shit", and "pandemic" from the word cloud.

The count of word in positive and negative dataset using unigram analysis were also identified. The unigram is the single word representation in the dataset. Fig. 7 shows the positive dataset's top 10 words: new, normal, home, social, work, mask, get, distancing, wear, and hand. The most used phrase is "new". While Fig. 8 shows the top 10 words in the negative dataset: home, work, mask, get, normal, new, people and hand. The words in both unigrams mostly are the same, but the count for all words is different. For example, the term "new" in the positive Data Frame is more than 400, while the word "new" on the negative side is less than 100.

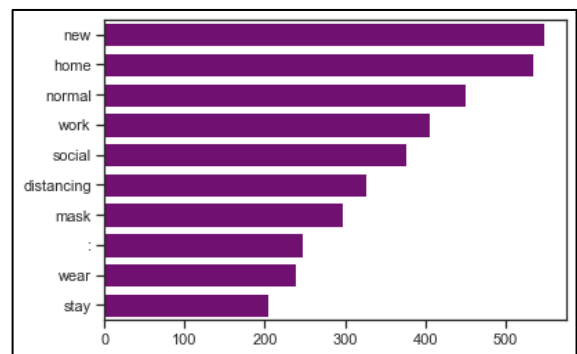


Fig. 4. Top 10 Common Words.



Fig. 5. Word cloud for Positive Datasets.

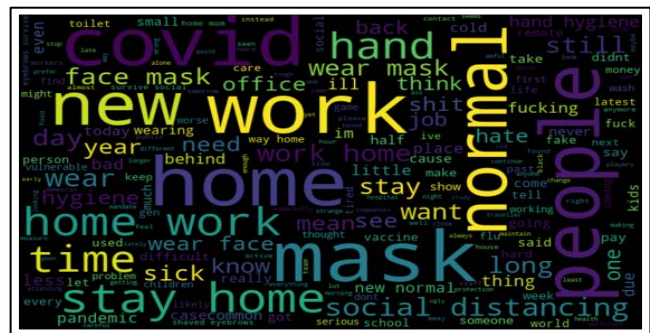


Fig. 6. Word cloud for Negative Dataset.

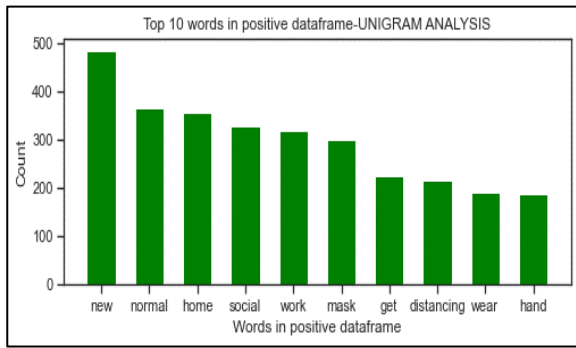


Fig. 7. Unigram Analysis for Words in Positive Dataset.

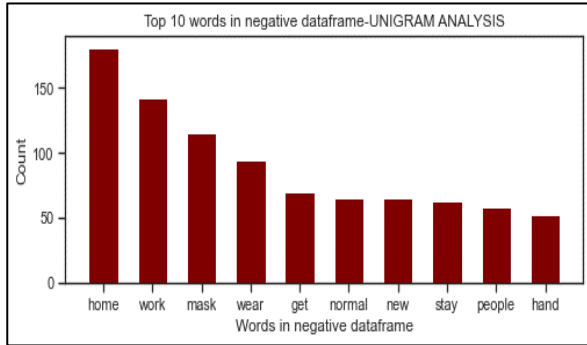


Fig. 8. Unigram Analysis for Words in Negative Dataset.

### B. Naïve Bayes Classifier Performance Evaluation

This section discussed the performance of the Naïve bayes classifier. The first performance was evaluated by looking at the accuracy of the classifier to classify the sentiment of the tweet data. It is done by comparing the actual label in testing data with the predicted label provided by the classifier. In this study, the accuracy was calculated by using the sklearn library in python. Table V presents the accuracy results. Based on the testing, the best accuracy is 83%, and the minimum is 79%.

To better understand the classifier performance, a confusion matrix was used. The analysis is focused on the 90:10 dataset, which obtained the best accuracy result. In this dataset, 281 data were tested. The confusion matrix obtained is shown in Fig. 9. The figure shows that there are 180 positive data and 54 negative data that the model has correctly predicted. However, the Naïve Bayes model cannot predict the remaining 47 data.

The confusion matrix can calculate the accuracy, precision, F1-score and recall of the Naïve Bayes model. Fig. 10 shows the detailed result of the classifier performance. The weighted average for precision, recall and F1- Score is 0.85, 0.83 and 0.84, respectively. Based on these values, even though the dataset contains an imbalance number of positive and negative data, the Naïve Bayes classifier can do the classification with 84% as indicated by the F1-score. In addition, all the parameters show a consistent value, such as the accuracy of the model.

Fig. 11 shows the ROC curve for the Naïve Bayes model. The true positive rate (TPR) is plotted against the false positive rate (FPR) to create a ROC curve (FPR). The actual positive rate (TP/ (TP + FN)) is the proportion of positive

observations that were correctly expected to be positive out of all positive observations. The closer the ROC curve approaches the upper left corner of the plot, the more effectively the model classifies data. To determine how much of the plot falls under the curve, the AUC (area under the curve) is used. The AUC value for Naïve Bayes model is 0.82. The greater the AUC, the better the model's ability to distinguish between the positive and negative classes of data [28].

TABLE V. ACCURACY OF THE CLASSIFIER FOR EACH DATASET

Split dataset	Train data	Test data	Accuracy
90:10	2526	281	83%
80:20	2245	562	81%
70:30	1964	843	79%

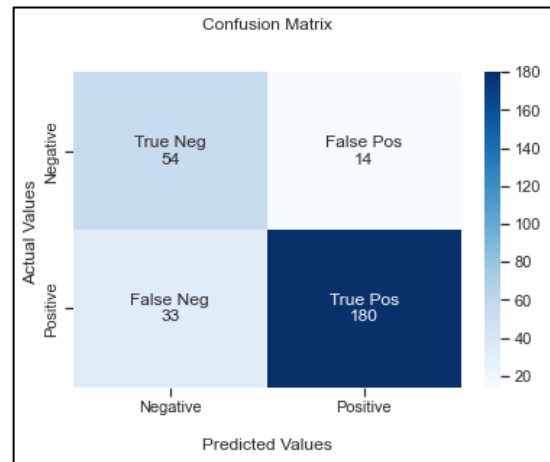


Fig. 9. Confusion Matrix for 90:10 Split Dataset.

	precision	recall	f1-score	support
-1	0.62	0.79	0.70	68
1	0.93	0.85	0.88	213
accuracy			0.83	281
macro avg	0.77	0.82	0.79	281
weighted avg	0.85	0.83	0.84	281

Fig. 10. Classification Report.

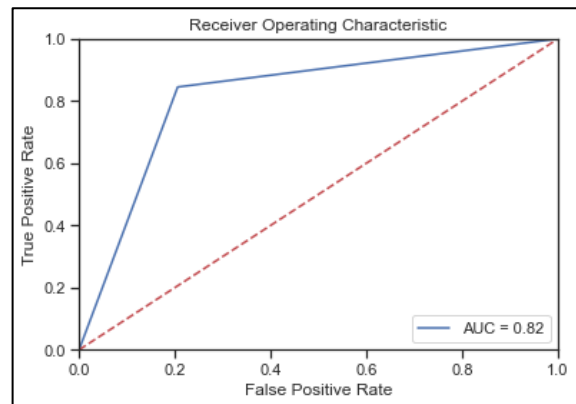


Fig. 11. ROC Curve.



This study also compares the accuracies of the Naïve Bayes algorithm implemented in other similar works. Table VI shows the accuracies of Naïve Bayes in each of the studies. Based on Table VI, the Naïve Bayes algorithm has generated good accuracies of more than 80% in all similar works. A Naïve Bayes algorithm could produce good and acceptable performance in sentiment classification problems. It is proven that Naïve Bayes is a reliable classifier due to its capabilities in solving various classification problems.

TABLE VI. COMPARISON OF NAÏVE BAYES ACCURACY BETWEEN SIMILAR WORKS

Authors	Accuracy
[3]	80.37%
[9]	80.65%
[29]	84.1%
Proposed Naïve Bayes Classifier	83%

### C. The Proposed Prototype

A prototype has been proposed for the implementation of the classifier. Fig. 12 displays the main user interface for the sentiment analyzer system. The user prototype was developed using the Python library's Tkinter framework. In the system, the user needs to input the sentiments first and then click on the "Check Sentiment Result" button to obtain the sentiment results. The result will then show the category of the tweets, whether it is Positive or Negative. The user needs to click on the "Clear" button to analyze the following statements. In this initial development, the system could only process one tweet at a time. The next improvement would enable the system to process several tweets simultaneously.

Fig. 13 shows the model's accuracy interface. If the user wants to see the model's accuracy in predicting the sentiment, the user must click the "Check Accuracy" button. It will navigate the user to a new window that shows the accuracy. The "Exit" button is used to close the system.



Fig. 12. Main Interface.

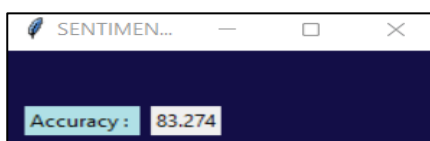


Fig. 13. Model's Accuracy Interface.

### D. Research Limitation

Several limitations of this study have been identified during the project's development. The first one is the quantity of data that has been scrapped is quite small. It is due to the time constraints of the project. Also, the standard API only allowed to retrieve tweets up to seven days and has a limited number to retrieve the data. This is because of the restriction on the Twitter developer account. In addition, the distribution of negative and positive tweets is also unbalanced. This might affect the performance of the classifier [30].

The next limitation is emotion analysis, which is not considered in this project. This project cannot indicate the public's emotion toward the new normal issues in the COVID-19 pandemic. The emotions are such as the people feel angry, surprised, happy or sad about accepting the new normal in their daily life. Lastly, the scrapped tweets are limited to English tweets and are not filtered by specific locations. The insight could not be generalizable to non-English speaking populations if only English tweets are used as the dataset. In addition, since most tweets do not have geolocation, it could be lacking in making conclusions based on certain countries or regions [31].

### V. CONCLUSION

This study has successfully explored the capability of the Naive Bayes algorithm in solving the sentiment classification on the acceptance of new normal in the COVID-19 pandemic. A total of 2807 tweets have been processed, which consisted of 2095 positive and 712 negative tweets. Based on the evaluation results, Naive Bayes has generated good and acceptable performance with 83% accuracy and 84% of F1-score. In addition, the developed Naïve Bayes classifier can distinguish between positive and negative tweets as indicated by AUC value of 0.82.

The significance of this study is in demonstrating the capability of the Naïve Bayes classifier in sentiment analysis. The proposed conceptual framework shown in Fig. 3 can be used as a guideline in conducting similar works. As for the study on the acceptance of the new normal in the COVID-19 pandemic, exploratory data analysis on the tweets showed more positive sentiments than negative ones. This indicates that most people could accept the new normal in their daily life during the COVID-19 pandemic. The government could use the results of this study as a resource for consideration in developing policies and campaigns and making approaches to the people to implement the new normal. This study could help the ministry of health deliver the necessary messages to the public while also addressing public concerns and encouraging positive behaviour in response to the COVID-19 pandemic.

Future works would be to include the public's emotions and to process non-English tweets. The next scrapped Twitter data would also be based on geolocation, so that data could be analyzed based on particular countries or regions.

### ACKNOWLEDGMENT

Special gratitude goes to Universiti Teknologi MARA Cawangan Terengganu for the continuous support given



towards the completion of this project. This research is funded under the UiTM Geran Penyelidikan Myra Lepas PHD (600-RMC/GPM LPHD 5/3 (068/2021)).

#### REFERENCES

- [1] Pozzi, E. Fersini, E. Messina, and B. Liu, "Challenges of sentiment analysis in social networks: an overview," *Sentiment analysis in social networks*, pp. 1-11, 2017.
- [2] M. Wongkar and A. Angdresey, "Sentiment analysis using Naive Bayes Algorithm of the data crawler: Twitter," in *2019 Fourth International Conference on Informatics and Computing (ICIC)*, 2019: IEEE, pp. 1-5.
- [3] S. Kaparang, D. R. Kaparang, and V. P. Rantung, "Analisis Sentimen New Normal Pada Masa Covid-19 Menggunakan Algoritma Naive Bayes Classifier," *Jointer - Journal of Informatics Engineering*, 2021.
- [4] W. H. Organization, "COVID-19 The New Normal | WHO Malaysia," [www.who.int](https://www.who.int), 2022. [Online]. Available: <https://www.who.int/malaysia/emergencies/covid-19-in-malaysia/information/the-new-normal>.
- [5] F. Lendriyono, "Public's Perception on Social Media towards New Normal during Covid-19 Pandemic in Indonesia: Content Analysis on Religious Social Media Accounts," in *IOP Conference Series: Earth and Environmental Science*, 2021, vol. 717, no. 1: IOP Publishing, p. 012039.
- [6] A. Cattapan, J. M. Acker-Verney, A. Dobrowsky, T. Findlay, and A. Mandrona, "Community engagement in a time of confinement," *Canadian Public Policy*, vol. 46, no. S3, pp. S287-S299, 2020.
- [7] J. Kaos, "PM: M'sia will transition into endemic phase from April 1," *The Star*, 2022. [Online]. Available: <https://www.thestar.com.my/news/nation/2022/03/08/pm-msia-will-enter-endemic-phase-from-april-1>.
- [8] E. Azeraf, E. Monfrini, and W. Pieczynski, "Improving usual Naive Bayes classifier performances with Neural Naive Bayes based models," *arXiv:2111.07307 [cs, stat]*, 2021, doi: <https://doi.org/10.48550/arXiv.2111.07307>.
- [9] T. B. T. Dewi, N. A. Indrawan, I. Budi, A. B. Santoso, and P. K. Putra, "Community Understanding of the Importance of Social Distancing Using Sentiment Analysis in Twitter," in *2020 3rd International Conference on Computer and Informatics Engineering (IC2IE)*, 2020: IEEE, pp. 336-341.
- [10] C. Shofiya and S. Abidi, "Sentiment analysis on COVID-19-related social distancing in Canada using Twitter data," *International Journal of Environmental Research and Public Health*, vol. 18, no. 11, p. 5993, 2021.
- [11] M. H. Tsai and Y. Wang, "Analyzing Twitter data to evaluate people's attitudes towards public health policies and events in the era of COVID-19," *International Journal of Environmental Research and Public Health*, vol. 18, no. 12, p. 6272, 2021.
- [12] S. N. Saleh, C. U. Lehmann, S. A. McDonald, M. A. Basit, and R. J. Medford, "Understanding public perception of coronavirus disease 2019 (COVID-19) social distancing on Twitter," *Infection Control & Hospital Epidemiology*, vol. 42, no. 2, pp. 131-138, 2021.
- [13] H. T. Zaw, N. Maneerat, and K. Y. Win, "Brain tumor detection based on Naive Bayes Classification," in *2019 5th International Conference on engineering, applied sciences and technology (ICEAST)*, 2019: IEEE, pp. 1-4.
- [14] S. Singleton, S. A. P. Kumar, and Z. Li, "Twitter Analytics-Based Assessment: Are the United States Coastal Regions Prepared for Climate Change?," presented at the *2018 IEEE International Symposium on Technology and Society (ISTAS)*, Washington DC, DC, USA, 2018. [Online]. Available: <https://doi.org/10.1109/ISTAS.2018.8638266>.
- [15] C. Bemando, E. Miranda, and M. Aryuni, "Machine-Learning-Based Prediction Models of Coronary Heart Disease Using Naive Bayes and Random Forest Algorithms," in *2021 International Conference on Software Engineering & Computer Systems and 4th International Conference on Computational Science and Information Management (ICSECS-ICOCSIM)*, 2021: IEEE, pp. 232-237.
- [16] P. Vadapalli, "Naive Bayes Classifier: Pros & Cons, Applications & Types Explained," *upGrad blog*, 2020. [Online]. Available: <https://www.upgrad.com/blog/naive-bayes-classifier/#:~:text=to>.
- [17] F. Razaque et al., "Using naïve bayes algorithm to students' bachelor academic performances analysis," in *2017 4th IEEE International Conference on Engineering Technologies and Applied Sciences (ICETAS)*, 29 Nov.-1 Dec. 2017 2017, pp. 1-5, doi: 10.1109/ICETAS.2017.8277884.
- [18] V. Radpour and F. S. Gharehchopogh, "A Novel Hybrid Binary Farmland Fertility Algorithm with Naive Bayes for Diagnosis of Heart Disease," *Sakarya University Journal of Computer and Information Sciences*, vol. 5, no. 1, pp. 90-103, 2022.
- [19] R. Setik, R. M. T. R. L. Ahmad, and S. Marjudi, "Exploring Classification For Sentiment Analysis From Halal Based Tweets," in *2021 2nd International Conference on Artificial Intelligence and Data Sciences (AiDAS)*, 8-9 Sept. 2021 2021, pp. 1-6, doi: 10.1109/AiDAS53897.2021.9574255.
- [20] J. Brownlee, "A Gentle Introduction to the Bag-of-Words Model," *Machine Learning Mastery*, 2019. [Online]. Available: <https://machinelearningmastery.com/gentle-introduction-bag-words-model/>.
- [21] Kiilu, Okeyo, Rimiru, and Ogada, "Using Naive Bayes Algorithm in detection of Hate Tweets," [www.ijsrp.org](http://www.ijsrp.org), 2018. [Online]. Available: <https://www.ijsrp.org/research-paper-0318.php?rp=P757259>.
- [22] H. Kamel, D. A. Abdulah, and J. M. Al-Tuwaijari, "Cancer Classification Using Gaussian Naive Bayes Algorithm," *2019 International Engineering Conference (IEC)*, pp. 165-170, 2019.
- [23] R. L. Mustafa and B. Prasetyo, "Sentiment analysis using lexicon-based method with naive bayes classifier algorithm on #newnormal hashtag in twitter," *Journal of Physics: Conference Series*, vol. 1918, no. 4, p. 042155, 2021/06/01 2021, doi: 10.1088/1742-6596/1918/4/042155.
- [24] P. Huilgol, "BoW Model and TF-IDF For Creating Feature From Text," *Analytics Vidhya*, 2020. [Online]. Available: <https://www.analyticsvidhya.com/blog/2020/02/quick-introduction-bag-of-words-bow-tf-idf/>.
- [25] B. Priyoko and A. Yaqin, "Implementation of Naive Bayes Algorithm for Spam Comments Classification on Instagram," *2019 International Conference on Information and Communications Technology (ICOIACT)*, pp. 508-513, 2019.
- [26] S. Narkhede, "Understanding AUC - ROC Curve," *Medium*, 2018. [Online]. Available: <https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5>.
- [27] R. Bhargava, "Word Clouds for Fun and Qualitative Data Analysis," *Medium*, 2017. [Online]. Available: <https://medium.com/@rahulbot/word-clouds-for-fun-and-qualitative-data-analysis-c81ea0c53868>.
- [28] A. Bhandari, "AUC-ROC Curve in Machine Learning Clearly Explained," *Analytics Vidhya*, 2020. [Online]. Available: <https://www.analyticsvidhya.com/blog/2020/06/auc-roc-curve-machine-learning/>.
- [29] A. R. Isnain, N. S. Marga, and D. Alita, "Sentiment Analysis Of Government Policy On Corona Case Using Naive Bayes Algorithm," *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, 2021.
- [30] Y. Li, H. Guo, Q. Zhang, M. Gu, and J. Yang, "Imbalanced text sentiment classification using universal and domain-specific knowledge," *Knowledge-Based Systems*, vol. 160, pp. 1-15, 2018.
- [31] P. Zola, C. Ragno, and P. Cortez, "A Google Trends spatial clustering approach for a worldwide Twitter user geolocation," *Information Processing & Management*, vol. 57, no. 6, p. 102312, 2020.

# Partial Differential Equation (PDE) based Hybrid Diffusion Filters for Enhancing Noise Performance of Point of Care Ultrasound (POCUS) Images

Deepa V S<sup>1</sup>, Jagathyraj V P<sup>2</sup>, Gopikakumari R<sup>3</sup>

College of Engineering Chengannur, Kerala, India<sup>1</sup>

Cochin University of Science and Technology, Kochi, Kerala, India<sup>1, 2, 3</sup>

**Abstract**—A hybrid filter is developed by combining smoothing and edge preservation properties of anisotropic diffusion (AD) filters and noise reduction features of median filtering. Mixed Gaussian Impulse noise and speckle noise are considered for analysis. The performance of this hybrid filter is verified using ultrasound images. The effectiveness of this filter is assessed with Point of Care Ultrasound (POCUS) images to verify whether the algorithm developed is applicable to them. POCUS refers to a handheld portable ultrasound instrument that can be used at patient bedside. Quantitative analysis with COVID-19 POCUS images, in terms of SNR, SSIM and MSE is performed. Results demonstrate that for all test images, the proposed filter has the best SNR, least MSE, and highest SSIM. Significant improvement in image quality is thus observed both qualitatively and quantitatively. The novelty of suggested technique is its effectiveness in reducing both mixed Gaussian impulse noise and speckle noise in ultrasound as well as POCUS images without the need for separate filters. POCUS has played a significant role in the diagnosis and management of pulmonary, cardiac and vascular pathologies associated with COVID-19. Automatic segmentation of these images and subsequent automatic detection and diagnosis are becoming increasingly popular due to the rapid development of artificial intelligence technologies. These results are useful in implementing better pre-processing prior to segmentation of ultrasound images to facilitate improved patient care.

**Keywords**—Anisotropic diffusion filter; POCUS; mixed Gaussian impulse noise; speckle noise

## I. INTRODUCTION

The use of ultrasound imaging for medical diagnosis is widely accepted due to its non-invasiveness, no risk factor, and efficiency. However, the acquisition process introduces noise in the signal, which has an impact on subsequent processes like segmentation, quantitative analysis, etc. The granular interference called, speckle noise is inherent in Ultrasound images. Impulse noises are yet another sort of noise present in ultrasound imaging. Another common type of noise found in medical images is Additive White Gaussian noise. Several different kinds of filters must be designed in order to effectively remove these noises. A median filter is a good choice for eliminating impulsive type noises. But it cannot suppress median tailed noise distributions like Gaussian. If linear filters are used to process such noise, they tend to blur the edges of the image [1]. Also, they cannot remove mixed Gaussian impulse noise and speckle noise adequately [2].

Studies show that various types of nonlinear filters [3] can be effectively used to remove such noises. Partial Differential equation based Anisotropic Diffusion (AD) filters are known for their ability to preserve edges in an image during denoising. AD approaches are being used in image processing since 1987 when Perona and Malik [4] introduced a non-linear method of edge preserving smoothing that outperformed the existing traditional linear methods [5]. Since ultrasound images are mostly affected by speckle noise and impulse noises [6], a combination of filter structures, which can filter out all types of noises, need to be derived. Median filter is usually employed for suppressing impulse noises like salt and pepper noise. However, it is not effective for reducing Gaussian noise or speckle noise [4]. Anisotropic Diffusion filters are the best choice in removing Gaussian noise and speckle noise. This paper introduces a new hybrid form of median and AD filters combining the advantages of median filters in removing impulse noise and AD filters in rejection of Gaussian and speckle noise. This hybrid form is found adequate for the removal of both mixed Gaussian impulse noise and speckle noise. The results are verified qualitatively and quantitatively.

In 1987, Perona and Malik proposed Nonlinear Anisotropic diffusion [4]. It is a filtering technique based on partial differential equation (PDE). It performs nonlinear smoothing and effectively reduces the image noise. The salient feature of AD filtering is that it can preserve important image features such as edges. While smoothing the rest of the image, it can maintain crisp texture detail at all viewing orientations [7]. It implies that blurring of edges and thus loss of information can be avoided. [8] Provides a derivation of AD filters for speckle reduction. Speckle noise is a form of multiplicative noise, usually present in medical ultrasound images and Synthetic aperture radar (SAR) images. Nonlinear means of removing speckle noise is vital in such cases. Some of the classical speckle removing filters like Lee filter or Frost filter tends to remove some important data also. Recent developments based on anisotropic diffusion filtering overcome the major drawbacks of conventional spatial filtering [8][9], and significantly improve image quality and provide better results than above mentioned filters[9][10].

Motivated from the work of Perona and Malik, various additive as well as multiplicative noise removal algorithms have been developed. A speckle reducing anisotropic diffusion (SRAD) method was proposed by Yu and Acton [8] which handles various noise distributions, especially, speckle. Further

improvement of the SRAD was presented by Karl et al. with the oriented speckle reducing anisotropic diffusion (OSRAD) method [11], incorporating local directional variance of image intensity. Both these methods have the drawback of producing over smoothed images. This problem was solved by anisotropic diffusion with memory based on speckle statistics (ADMSS) method [12] by Ramos, Zhou et al. [13] proposed a doubly degenerate nonlinear diffusion (DDND) model by using the diffusion equation theory. It guides the denoising process with the aid of the gradient information and the grey level information. In [14] speckle noise suppression and image segmentation of ultrasound image using AD filters with an improved diffusion coefficient is discussed. In [15] the drawbacks of SRAD filter are eliminated using an optimization algorithm for diffusion coefficient. The algorithm well removes speckle and is more suitable for image segmentation. However, no other noise than speckle is considered [16][17][18]. In [21], Mei Gao et al. proposed a filtering scheme for ultrasound images, which the noise at the edge is processed during denoising process. This is achieved by analyzing the divergence term. Most of the above mentioned developments and research, aided in speckle noise removal of ultrasound images. Such an extensive research has been done in the area of AD filters viewing its selective smoothing and edge preserving capability and speckle denoising property [19][20][22][23][24]. But almost all of them works on speckle removal only and doesn't mention about the other relevant noises.

However, in addition to speckle noise, impulse noise is present in ultrasound images and Gaussian noise is common in medical images. None of the above works consider the removal of such noises. In [2], Meenavati and Rajesh proposed a method to remove mixed Gaussian impulse noise from images using volterra filters. But the analysis does not consider ultrasound images or reduction of speckle noise. Since speckle is an important consideration in US images, AD with removal of speckle as well as mixed Gaussian impulse noise is significant. All these discussions clearly demand the development of a filter which can eliminate both speckle and mixed Gaussian impulse noise.

The uniqueness of proposed work is that the same filter can be used for reduction of both mixed Gaussian impulse noise and speckle noise. In this paper, the emphasis is given to the analysis of POCUS (Point of Care Ultrasound) images to verify whether the algorithm developed is applicable to POCUS images. Analysis with parameters like SNR, MSE, and SSIM is done to quantitatively verify the performance of the proposed filter for POCUS images. Results are found to be better than using median and simple AD filters for noise removal.

Significances of the work are listed below:

- Addresses the removal of almost all kinds of noise such as mixed Gaussian impulse noise and speckle noise, whereas the previous literature on AD primarily discusses speckle noise.

- No prior work has considered the denoising of POCUS images.
- Qualitative and quantitative analysis gave better results with high PSNR, least MSE and improved SSIM compared to the existing methodologies.

This paper is organized as follows. In Section II, a brief review of POCUS is given. The design features of median and nonlinear AD filters are analyzed in Section III. In Section IV, the features of the proposed filter are discussed. Methodology of work is presented in Section V. Experimental results and Quantitative analysis using these images is given in Section VI. Concluding remarks are presented in Section VII.

## II. A REVIEW OF POCUS

Point-of-care ultrasonography (POCUS) refers to handheld portable ultrasound instrument that can be used at patient bedside. In the midst of COVID-19 pandemic, such hand-carried ultrasound devices emerge as a tool that can simplify the imaging process [25]. These devices are perfect for COVID-19 scans because they are small enough to be covered completely with a probe cover and due to its small size, the decontamination process is also simplified. Several studies have indicated that wrapping the whole device in plastic or using single-use plastic sterile probe covers is enough to condense the decontamination process. In addition, a trained healthcare provider requires only 5 to 10 minutes to conduct a lung POCUS study [25]. Experts from China have specifically advocated for the use of hand-held POCUS in COVID-19 due its clinical and economic value [26]. The utility of these devices in continuous monitoring COVID-19 patients managed at home have also been reported. The salient features of POCUS such as its ability to connect to smartphones and tablets, artificial intelligence-assisted diagnosis, wireless feature, rechargeable batteries, and low cost make them a convenient and practical imaging option suitable even in remote areas [25][26][27]. Research has shown that point-of-care ultrasound device can help manage infectious diseases, as well as abdominal cardiac and pulmonary pathologies [25]-[29]. Images of some of the POCUS instruments available in market are shown Fig. 1



(a) Lumify Portable Ultrasound by Philips

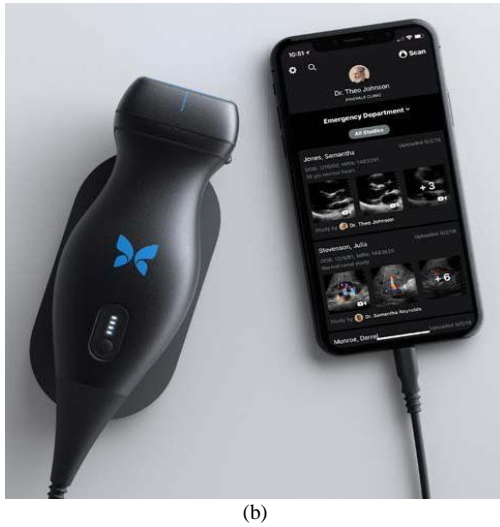


Fig. 1. (a) Lumify Portable Ultrasound by Philips (b) The Breakthrough Butterfly iQ Vet ultrasound System (Photo Courtesy of Butterfly Network, Inc.).

### III. FILTERING TECHNIQUES

#### A. Anisotropic Diffusion Filters

As presented by Perona and Malik in [4], the basic PDE equation of anisotropic diffusion can be represented as

$$\frac{\partial I}{\partial t} = \text{div}(c(x, y, t)\nabla I) = \nabla c \nabla I + c(x, y, t)\Delta I \quad (1)$$

Here original Image is  $I_0(x, y)$ .  $I(x, y, t)$  is the smoothed image via anisotropic diffusion method as the solution of equation (1).  $\Delta$  is the Laplacian operation,  $\nabla$  is the gradient of the image.  $\text{div}(\dots)$  denotes the divergence operator and  $c(x, y, t)$  is the diffusion coefficient.  $c(x, y, t)$  is a function of the image gradient which preserves edges and controls filtering process by controlling the rate of diffusion. The Diffusion coefficient can be evaluated by the two functions:

$$c(\|\nabla I\|) = e^{-\left(\frac{\|\nabla I\|}{k}\right)^2} \quad (2)$$

$$c(\|\nabla I\|) = \frac{1}{1 + \left(\frac{\|\nabla I\|}{k}\right)^2} \quad (3)$$

where  $k$  is the edge magnitude parameter.

A four-neighbourhood discrete form of (1) is given by

$$I(x, y, t + \Delta t) = I(x, y, t) + \frac{\Delta t}{4 \sum_{\rho \in Z} G(\nabla I(\rho, t))} \quad (4)$$

where  $Z$  is the set of the four neighbourhoods of pixel  $(x, y)$ , denotes a neighbourhood of  $(x, y)$ , and  $\nabla I(\rho, t) = I(\rho, t) - I(x, y, t)$  is the image gradient at current time  $t$ . The above equation is recursive over time until it meets the stopping criterion. Perona and Malik suggested that a desirable diffusion coefficient should satisfy the basic condition that it diffuses more in smooth areas and less around high-intensity transitions. By this technique, noise or unwanted texture is smoothed, while edges are sharpened [30]. The function  $G$  is a monotonically decreasing function, the edge magnitude parameter. Depending upon the value of the monotonically decreasing function  $G$  and the  $|d|$  which is the absolute value of

gradient, the anisotropic diffusion filtering can be formulated as follows:

- The range of  $G$  is  $[0, 1]$ . For any given parameter  $k$ , Monotonically decreases with  $|d|$ . If  $|d| \rightarrow 0$  then  $G \rightarrow 1$  is isotropic diffusion (Gaussian filtering); if  $|d| \rightarrow \infty$  then  $G \rightarrow 0$  the diffusion flow is arrested and the edges are preserved.
- For any given  $|d|$ ,  $G$  monotonically increases with parameter  $k$ , which means that  $k$  controls the generosity of the anisotropic diffusion filter. For higher value of  $k$ , the diffusion process is more likely to smooth the image and reduce the noise; while for lower value of  $k$ , the diffusion process is more restricted and is more likely to preserve image features [31].

The advantage of this technique is that it reduces noise and preserves the edges so that crisp edge features will be obtained.

#### B. Median Filter

Median filter is a nonlinear filter used for noise reduction in images. Each pixel value is obtained by taking the median value of neighboring pixels under the window. Thus, the result is the middle value after the input values have been sorted. When an image is considered, each pixel of the filtered image is replaced by median brightness value of its neighbourhood pixels in the original image.

Median filtering is a kind of smoothing technique like Gaussian filtering. Almost all the smoothing techniques including Gaussian filter adversely affect edges since they blur the image. Preserving edges is critically important for visual appearance of the image. For moderate levels of Gaussian noise (medium tailed distribution), median filter performs better than Gaussian filter and preserve edges. However, its performance is not significantly improved for high noise levels [32]. For removing salt and pepper noise (impulsive noise) median filters are more effective.

#### C. Proposed Hybrid Filter

The proposed filter is developed by combining the ability of median filter to remove impulsive noises with capabilities of AD filters to filter out Gaussian noise and speckle noise. Designing process involves two steps. The first step is to determine the median value of each and every pixel in the image being analysed. In the second step, these median values are used to design discretised form of anisotropic diffusion equation given in (5). To incorporate advantages of median filter, each pixel of the noised image used in the anisotropic diffusion process is replaced by median value its neighbourhood pixels and further processing is done. As described in [9], the discretised form of Perona –Malik Anisotropic diffusion equation is

$$I_{t+1}(s) = I_t(s) + \frac{\lambda}{\eta_s} g_k(|\nabla I_{s,p}|) |\nabla I_{s,p}| \quad (5)$$

Where  $I$  is the discretely sampled image,  $s$  the pixel position in the 2D grid,  $t$  denotes iteration step,  $g$  denotes conductance function and  $k$  is the gradient threshold parameter [7]. Constant  $\lambda \in (0,1)$  determines the rate of diffusion and  $\eta_s$



denotes the spatial 4-pixel neighborhood of  $s$ ,  $\eta_s = \{ N,S,E,W \}$  are the neighboring pixels of  $s$  in North, South, East and West directions. Visualization of 2D discrete diffusion is given in Fig. 2. The degraded image with each pixel replaced by its median value is given to the AD filter so that better removal of noise can be achieved.

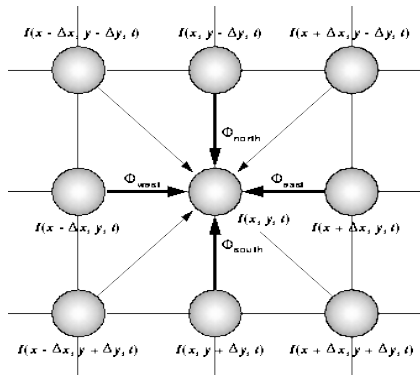


Fig. 2. Eight Neighborhood Visualisation of 2D Discrete Diffusion.

#### IV. METHODOLOGY

This work consists of four steps. They are

- Design of filters
- Implementation
- Qualitative Comparison
- Quantitative Comparison

Design is based on the design equations discussed in Section III. The work is implemented using Matlab R2018b.

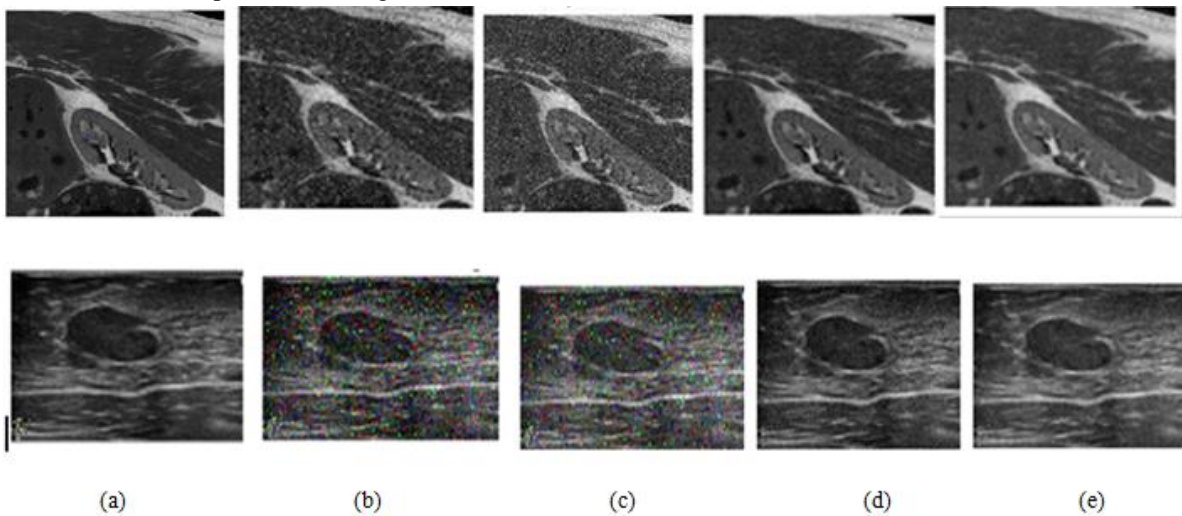


Fig. 3. Filter Response for Kidney Cut Ultrasound Image and Ben1 Image Corrupted by a Noise of Standard Deviation  $\sigma = 0.02$  and Impulse Noise Density  $\rho = 0.02$ . (a) Original Image (b) Mixed Gaussian Impulse Noised (c) Simple AD Filter (d) Median Filtered (e) Proposed Hybrid Filter.

#### A. Response of POCUS images to Mixed Gaussian Impulse Noise

POCUS images are collected from the dataset of [https://github.com/jannisborn/covid19\\_pocus\\_ultrasound](https://github.com/jannisborn/covid19_pocus_ultrasound) are used for the evaluation of noise filtering process. 50 POCUS images of COVID-19 are used from this dataset. Gaussian

The images used are degraded by mixed Gaussian impulse noise of standard deviation = 0.02 and impulse noise of density = 0.02 in order to assess how different filters respond to noise. Zero mean speckle noise with variance 0.04 is used for speckle noise analysis. In the proposed algorithm, the US image contaminated with mixed Gaussian impulse noise is decomposed into mask images considering 8-pixel neighborhood. The median filtered mask images are used to evaluate the diffusion coefficient in equation (3), which is further utilized for calculating AD algorithm. Initially, 15 iterations are done for simple AD and the proposed hybrid AD filters.

The qualitative analysis and comparison is performed by visually analyzing resultant images. By comparing SNR, MSE and SSIM, using the equations given in Section 4.2, quantitative analysis is done.

#### V. EXPERIMENTAL RESULTS

Kidney-cut ultrasound image of size 522X469 and Ben1 ultrasound image of size 538X317 are used for initial verification of noise filtering process. Gaussian noise of standard deviation  $\sigma = 0.02$  and impulse noise density  $\rho = 0.02$  are added to the image. The proposed hybrid filter output is compared with simple AD filter and median filter outputs. The output of kidney\_cut and Ben1 US images to simple AD filter, proposed hybrid filter and median filter are shown in Fig. 3. It can be seen that though simple AD filter is not good for eliminating Gaussian impulse noise, the performance is robust for the proposed hybrid AD filter. For median filter, visual quality seems comparatively far better than simple AD filter. But from quantitative analysis we can see that proposed hybrid filter outperforms median filter in SNR, MSE and SSIM.

noise of standard deviation  $\sigma = 0.02$  and impulse noise density  $\rho = 0.02$  are added to the image and the output is analysed. COVID-19 POCUS image, Cov\_severe, of size 367X367 is used for initial verification of noise filtering process. The proposed filter output is compared with simple AD filter and median filter. Results are shown in Fig. 4 and Fig. 5

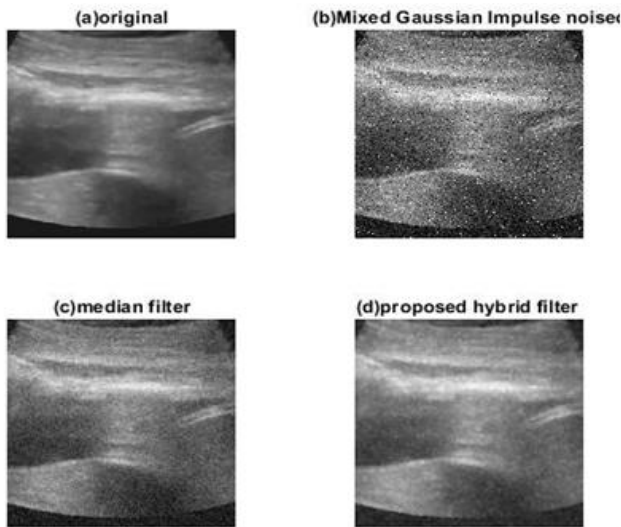


Fig. 4. Response of Various Filters (a) cov\_severe POCUS Image (b) Image Corrupted by Mixed Gaussian Impulse Noise (c) Response of Median Filter (d) Output of Proposed Filter.

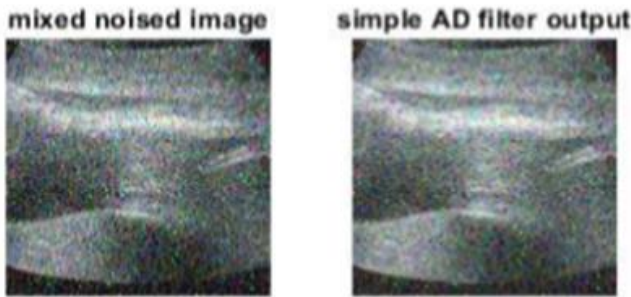


Fig. 5. Response of Simple AD Filter to cov\_severe POCUS Image.

Performance of the filter is evaluated using various POCUS images. A sample set of 10 images is shown in Fig. 6. Images used are degraded by mixed Gaussian impulse noise of standard deviation  $\sigma = 0.02$  and impulse noise of density  $\rho=0.02$

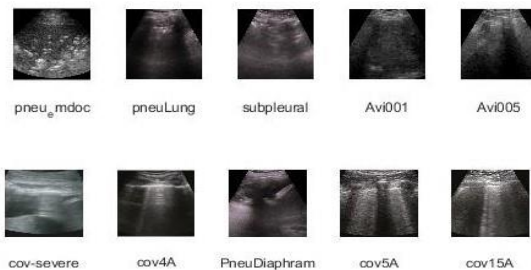


Fig. 6. Test Images used in Experiments.

Output for pneuLung image is shown in Fig. 7 and Fig. 8 respectively. For all the POCUS images, proposed filter provided better results. Visual quality of the proposed filter also seems better.



Fig. 7. Simple AD Filter (a) Original pneuLung POCUS Image (b) Image Corrupted by Mixed Gaussian Impulse Noise ( $\sigma=0.02$  and  $\rho=0.02$ ) (c) Response of Simple AD Filter.

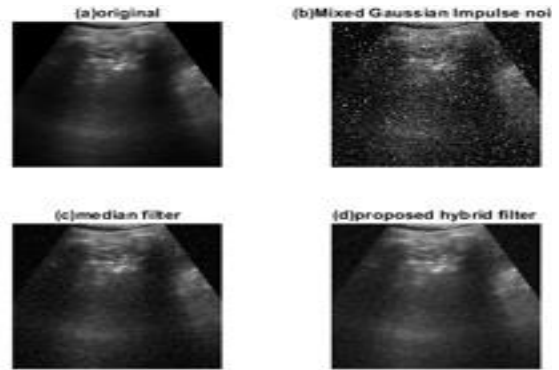


Fig. 8. Filter Response for pneuLung POCUS Image (a) Original Image (b) Mixed Gaussian Impulse Noised (c) Median Filter o/p (d) Proposed Filter Output.

### B. Response to Speckle Noise

The proposed hybrid AD filter performance is analyzed with an input corrupted by speckle noise. Results show that speckle removal can be efficiently achieved if the image is processed using the proposed hybrid filters.

Speckle noise is the major type of noise present in an ultrasound image. It limits contrast resolution of images by affecting the edges and fine details and make diagnostic more difficult. Anisotropic diffusion filters are efficient in removing speckle noise. Combining the advantages of median as well as AD filter, the performance of the proposed filter to POCUS images corrupted with zero mean speckle noise with variance 0.04 is analysed.

Fig. 9 compares outputs of simple AD filter, median filter and proposed hybrid filter responses for kidney\_cut US image for speckle noised image.

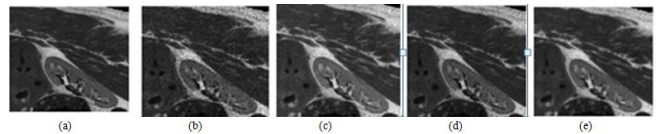


Fig. 9. Various Filter Responses for Kidney Cut Ultrasound Image for Speckle noise. (a)Original Image (b) Speckle Noised (c)Simple AD (d)Median Filter Output (e)Proposed Filter Output.

Fig. 10 and 11 compares outputs of simple AD filter, median filter and proposed hybrid filter responses for Cov\_severe POCUS image for speckle noised image. Here, the same zero means speckle noise with variance 0.04 is used. Here also, the visual quality is improved and noise removal is achieved.



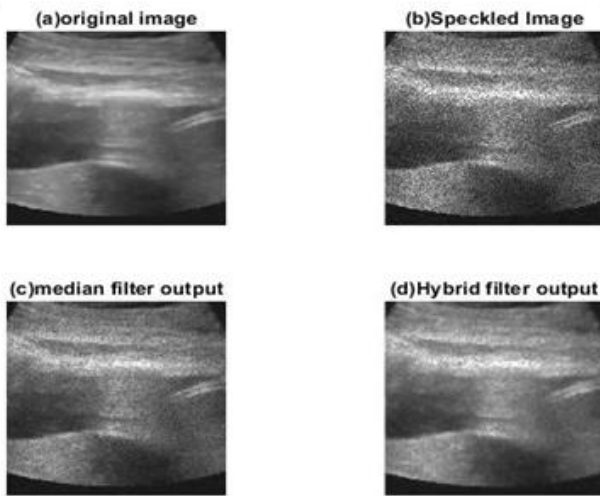


Fig. 10. Various Filter Responses for cov\_severe POCUS Image for Speckle Noise.

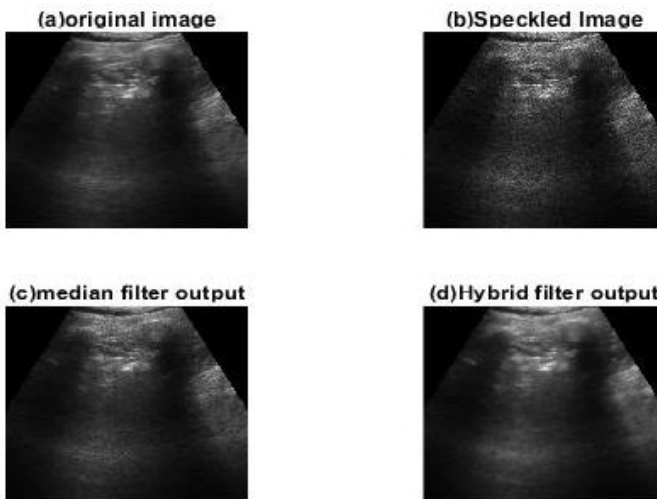


Fig. 11. Various Filter Responses for pneuLung POCUS Image for Speckle Noise. (a) Original Image (b) Speckle Noised (c) Median Filter o/p (d) Proposed Filter Output.

Simple AD filter response for pneuLung POCUS Image using zero mean speckle noise with variance 0.04 is shown in Fig. 11. For all these experiments, 15 iterations are done initially and verified the process using 20 and 30 iterations.

### C. Quantitative Analysis

A quantitative analysis is done to evaluate the performance of the proposed anisotropic diffusion filter by comparing the parameters such as Signal to Noise Ratio (SNR), Mean Square Error (MSE) [6] and structural Similarity Index (SSIM).

The SSIM is a method for measuring the similarity between two images. The SSIM index can be viewed as a quality measure of one of the images being compared while the other image is considered as of perfect quality. Maximum value of SSIM is 1, reachable only in the case of two identical sets of data [33].

The SNR and MSE are computed using the formula [2]:

$$SNR = 10 \log_{10} \frac{\sum_{p=1}^N \sum_{q=1}^N v(p,q)^2}{\sum_{p=1}^N \sum_{q=1}^N (u(p,q) - v(p,q))^2} \quad (6)$$

$$MSE = \frac{1}{N \times N} \sum_{p=1}^N \sum_{q=1}^N (u(p,q) - v(p,q))^2 \quad (7)$$

Quantitative analysis results with parameters SNR, MSE and SSIM for Ben1 and kidney\_cut US images are shown in Table I. Table II shows SNR, MSE and SSIM values for five sample COVID-19 Lung US images corrupted with mixed Gaussian impulse noise. The parameters were measured after 15 iterations using Simple AD filter, median filter and the proposed filtering method. Results are sketched in Fig. 12. A comparison of SSIM values are plotted in Fig. 13. From these sketches, it is clear that the proposed method outperforms the other methods and turns out to be the most robust scheme, as it yields better SNR, minimum MSE and highest SSIM for all the images.

TABLE I. QUANTITATIVE ANALYSIS OF BEN1 AND KIDNEY\_CUT US IMAGES ON MEDIAN, AD & PROPOSED FILTERS FOR MIXED GAUSSIAN IMPULSE NOISE

Image	Parameters	Mixed Gaussian Impulse noise		
		Median Filter	Simple AD filter	Proposed filter
Ben1	SNR (dB)	18	8	19.35
	MSE	28.16	31	2.96
	SSIM	0.52	0.25	0.65
Kidney cut	SNR (dB)	10.05	8	22.49
	MSE	25	32	1.4
	SSIM	0.52	0.19	0.69

TABLE II. QUANTITATIVE ANALYSIS OF POCUS IMAGES ON MEDIAN, AD & PROPOSED FILTERS FOR MIXED GAUSSIAN IMPULSE NOISE

Image	Parameter s	Mixed Gaussian Impulse noise		
		Median filter	Simple AD filter	Proposed filter
Cov_severe	SNR (dB)	10.5	8.7	25.9
	MSE	22.76	33.8	0.65
	SSIM	0.51	0.15	0.76
Pneu_lung	SNR (dB)	10.7	5.8	28.4
	MSE	18	19	0.36
	SSIM	0.4	0.1	0.6
Cov_5A	SNR (dB)	10.6	9	26.4
	MSE	22	30.8	0.58
	SSIM	0.59	0.2	0.74
Cov_15A	SNR (dB)	10	9.4	27
	MSE	21.8	30	0.5
	SSIM	0.53	0.15	0.73
Sub_pleu	SNR (dB)	10	8	27
	MSE	19.5	28	0.49
	SSIM	0.45	0.12	0.6

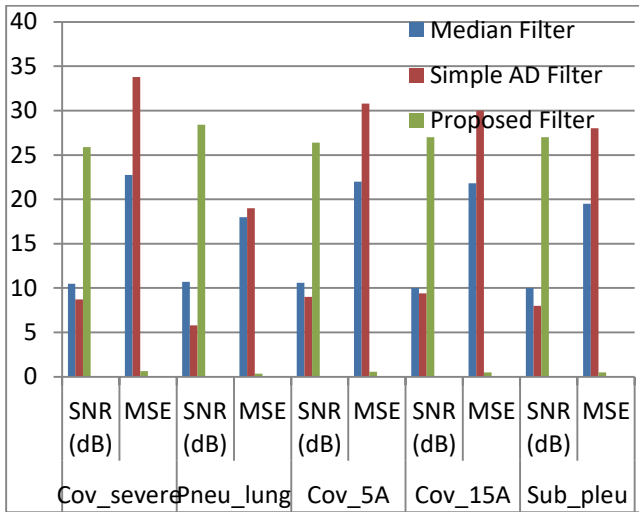


Fig. 12. Comparison of SNR, and MSE using 5 Sample Images with Mixed Gaussian Impulse Noise. Results of Median, Simple AD and Proposed Filtering Schemes.

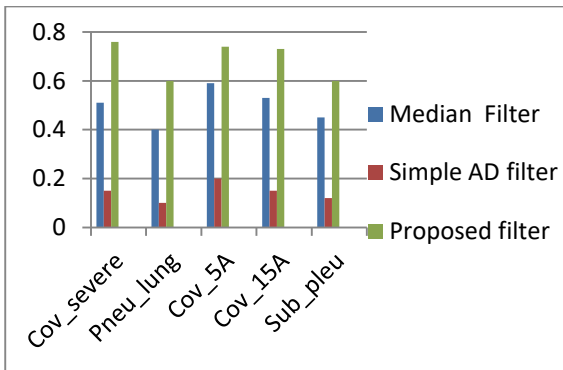


Fig. 13. Comparison of SSIM Values.

Table III shows the same parameters for images degraded with speckle noise. SNR and MSE plot is shown in Fig. 14 and SSIM plot in Fig. 15. From these plots, we can see that Simple AD filters are better in performance than median filters in speckle filtering process, while they perform poor in reducing Gaussian impulse noise. The proposed filter gives more satisfactory results for all the three parameters for all images with highest SNR value of 19.8, minimum MSE of 3 and maximum SSIM of 0.896.

TABLE III. QUANTITATIVE ANALYSIS OF PROPOSED FILTER, SIMPLE AD & MEDIAN FILTER FOR SPECKLE NOISE

Image	Parameters	Speckle Noise		
		Median Filter	Simple AD filter	Proposed filter
Cov_severe	SNR (dB)	7.5	13	19.2
	MSE	44	40	3
	SSIM	0.56	0.4	0.78
Avi_005	SNR	9.5	13	19.8
	MSE	28	5.5	2.6
	SSIM	0.73	0.82	0.84
Cov 15A	SNR (dB)	9.5	13	19
	MSE	28.28	19.9	2.7
	SSIM	0.7	0.63	0.84
Pneu_lung	SNR (dB)	12	15.8	17.5

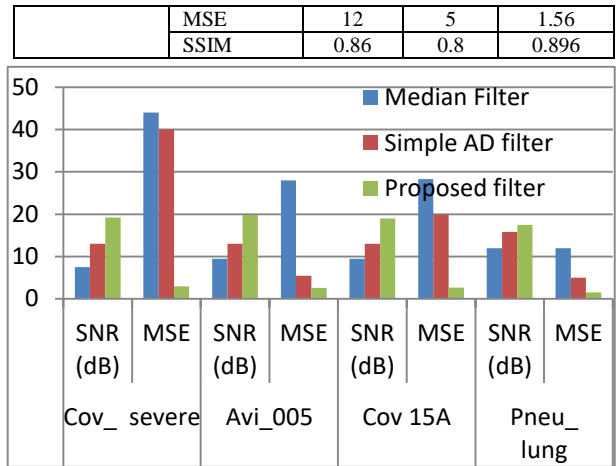


Fig. 14. Comparison of SNR and MSE using Four Sample Images with Speckle Noise. Results of Median, Simple AD and Proposed Filtering Schemes.

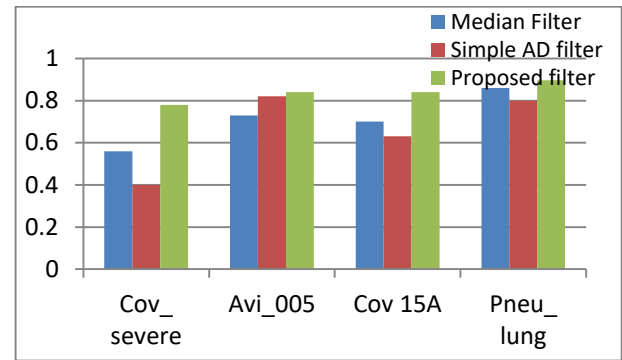


Fig. 15. Comparison of SSIM Values for Median, Simple AD and Proposed Filtering Schemes when Speckle Noise is Added to 4 Different Covid POCUS Images.

## VI. CONCLUSION

This paper introduces a new method of noise filtering of POCUS images based on hybrid anisotropic diffusion filters. The smoothing and edge preservation properties of AD filters and the noise reduction features of median filtering are combined to optimize the performance. Both mixed Gaussian Impulse noise and Speckle noise are considered. The resultant images are analysed quantitatively using parameters SNR, MSE, and SSIM. For mixed Gaussian impulse noise, the proposed filter yields a maximum SNR of 28.4 while median and simple AD filters gave only 10.7 and 9.4 respectively. Similarly, the proposed filter has the highest SNR value of 19.8 with speckle noise. The maximum MSE is 0.58, 30.8 and 21.8 for proposed, median and simple AD respectively, with proposed filter scoring minimum MSE for all images. With speckle noise, these MSE values are 3, 44 and 40 respectively. SSIM values are also the highest with 0.76 and 0.896 with mixed Gaussian impulse noise and speckle noise respectively. These findings show that the proposed filter delivers maximum SNR, least MSE, and highest SSIM for all test images. The results were uniform and consistent across all the test images after 20 and 30 iterations.

Due to its low cost, quick diagnosis, and non-exposure to radiation, Ultrasound is recommended in many clinical

scenarios, including respiratory, cardiovascular, and thromboembolic elements of COVID 19, obstetrics, etc. The development of artificial intelligence technology has made automatic segmentation and further diagnosis and detection excellent. A pre-processing stage prior to segmentation is inevitable in all these cases due to the presence of speckle and other noises, poor contrast, and acoustic shadows in US images. These robust software tools, when used in conjunction with point-of-care technologies, are well-suited to replace X-ray and CT scan on patient triage and immediate care.

#### REFERENCES

- [1] I. Pitas and A.N. Venetsanopoulos, "Edge detectors based on nonlinear filters," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 8, n. 4, pp. 538-550
- [2] M. B. Meenavathi, and K. Rajesh, "Volterra Filtering Techniques for Removal of Gaussian and Mixed Gaussian-Impulse Noise," *World Academy of Science, Engineering and Technology* 2, page. 238-244, 2007
- [3] Sanjit K. Mitra and Giovanni L. Sicuranza, *Nonlinear Image Processing* (NJ: Academic Press, 2001).
- [4] Perona and Malik (1990), "Scale space and edge detection using anisotropic diffusion," *IEEE Trans. Patt.Anal.Machine Intell*, 12:629–639.
- [5] Goyal, S., Rani, A., Yadav, N., & Singh, V., "SGSSRAD Filter for Denoising and Edge Preservation of Ultrasound Images," In 6th International Conference on Signal Processing and Integrated Networks (SPIN), (pp. 676-682), 2019.
- [6] Alka Vishwa, Shilpa Sharma, "Modified Method for Denoising The Ultrasound Images by Wavelet Thresholding," Published online in MECS, IJ Intelligent system and applications. June2012, Page: 25-30.
- [7] Weickert J, Snorr C, "PDE based Preprocessing of Medical Images," *Kunstliche Intell*, 14, page. 5-10, 2000.
- [8] Yongjian Yu and Scott T. Acton, Senior Member, "Speckle Reducing Anisotropic Diffusion," *IEEE Transactions on Image Processing*, VOL. 11, NO. 11, Nov 2002.
- [9] Choi H, Jeong J, "Despeckling Algorithm for Removing Speckle Noise from Ultrasound Images," *Symmetry*. 2020; 12(6):938.
- [10] Joachim Weickert, Department of Computer Science University of Copenhagen Copenhagen, Denmark, "Anisotropic Diffusion in image processing," B. G. Teubner (Stuttgart), 1998.
- [11] Krissian, K. Westin, C.F. Kikinis, R., Vosburgh, K.G. "Oriented speckle reducing anisotropic diffusion." *IEEE Trans. Image Process*. 2007, 16, 1412–1424.
- [12] Ramos-Llordén, G. Vegas-Sánchez-Ferrero, G. Martín-Fernández, M., Alberola-López, C.; Aja-Fernández, S. Anisotropic diffusion filter with memory based on speckle statistics for ultrasound images. *IEEE Trans. Image Process*. 2015, 345–358.
- [13] Zhou Z, Guo Z, Dong G, Sun J, Zhang, D, Wu B, "A doubly degenerate diffusion model based on the gray level indicator for multiplicative noise removal," *IEEE Trans. Image Process*. 2015, 249–260.
- [14] Vivian Bass, Julieta M, Ivan M. Rosado-Mendez, et al, "Ultrasound image segmentation methods: A review," *AIP Conference Proceedings* 2348, 050018 (2021)
- [15] Thakur, N., Khan, N.U. & Sharma, S.D, "A review on Performance analysis of PDE based Anisotropic Diffusion Approaches for Image Enhancement", *Informatica* 45 (2020), 89-102
- [16] B Balagalla, S Subasinghe, C de Alwis "A Review on ultrasound image pre-processing, segmentation and compression for enhanced image storage and transmission," *KDU International Research Conference*, sept 2018
- [17] Longsheng Wei, Dapeng Luo, Xeinmai Wang, "Active contour Texture image segmentation based on Anisotropic Diffusion," *Journal of Computational Information Systems*.
- [18] Tudor Barbu, "Robust Anisotropic Diffusion Scheme for Image Noise Removal," *Procedia Computer Science* (2014) 522 – 530.
- [19] Thakur, N., Khan, N.U. & Sharma, S.D, "An efficient fuzzy inference system based approximated anisotropic diffusion for image de-noising," *Cluster Comput* (2022).
- [20] Febin, I.P., Jidesh, P. "Despeckling and enhancement of ultrasound images using non-local variational framework," *Vis Comput* 38, 1413–1426 (2022).
- [21] Jidesh Pacheeripadikkal, A. Bini, "Image despeckling and deblurring via regularized complex diffusion", *Signal Image Video Processing*, pp. 977–984, 2017.
- [22] Yu Deng Hua Huang, "Ultrasound Image Segmentation Based on the Anisotropic Diffusion Filtering," ©2010 IEEE.
- [23] V J. Ge, Y. Dai, T. Chen, J. Zhang and X. Yao, "Ultrasound Image Filtering Using Partial Differential Equations," 2019 IEEE International Conference on Robotics and Biomimetics (ROBIO), 2019, pp. 1210-1215,
- [24] Removal Mei Gao, Baosheng Kang, Xiangchu Feng, Wei Zhang and Wenjuan Zhang, "Anisotropic Diffusion Based Multiplicative Speckle Noise," *Sensors* 2019, 19, 3164
- [25] Jannis Born, Gabriel Brändle, et al. "POCOVID-Net: Automatic Detection of COVID-19 From a New Lung Ultrasound Imaging Dataset (POCUS)," *ISMB TransMed COSI* 2020K.
- [26] Yau O, Gin K, Luong C, et al. "Point-of-care ultrasound in the COVID-19 era: A scoping review," *Echocardiography*. 2020;00:1–14.
- [27] Born J, Wiedemann N, Cossio M, Buhre C, Brändle G, Leidermann K, Aujayeb A, Moor M, Rieck B, Borgwardt K "Accelerating Detection of Lung Pathologies with Explainable Ultrasound Image Analysis," *Appl. Sci*. 2021, 11(2), 672. <https://doi.org/10.3390/app11020672>.
- [28] 1986X. Vandemergel, "Point-of-Care Ultrasound (POCUS) in the Field of Diabetology", *Hindawi International Journal of Chronic Diseases*, Volume 2021, Article ID 8857016, 8 pages
- [29] Smallwood N, Dachsel M, Smallwood N, et al. "Point-of-care ultrasound (POCUS): unnecessary gadgetry or evidence-based medicine". *Clin Med (Lond)*. 2018 Jun;18(3):219-224
- [30] Xiaoshuang Ma, Huanfeng Shen, Liangpei Zhang, Jie Yang, Hongyan Zhang. "Adaptive Anisotropic Diffusion Method for Polarimetric SAR Speckle Filtering", *IEEE Journal of Selected Topics in Applied Earth Observations and remote sensing*, vol. 8, no. 3, pp. 1041-1050, March 2015.
- [31] C. Tsotsios, M. Petrou, "On the choice of the parameters for anisotropic diffusion in image processing," *Pattern Recognition* (2012)
- [32] Arce, Gonzalo R. (2005). *Nonlinear Signal Processing: A Statistical Approach*. New Jersey, USA: Wiley.
- [33] Z. Wang, A.C. Bovik, H.R. Sheikh, E.P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing* 13 (4) (2004) 600–612

# Smart Greenhouse Monitoring and Controlling based on NodeMCU

Yajie Liu<sup>1,2</sup>

Post Graduate Center

Management and Science University, Shah Alam, Malaysia<sup>1</sup>

School of Information and Engineering,

Henan Vocational University of Science and Technology, China<sup>2</sup>

**Abstract**—Food security is one of the major rising issues as the human population is larger and the land available for cultivation is smaller, as well unassured affairs happened often in society especially in the current CoVID-19 rapidly spread days. To mitigate this condition, further improve the yields and quality of food, this paper proposed a smart and low-cost greenhouse monitoring and control system, which mainly consists of sensors, actuators, LCD display and microcontrollers. DHT22 sensor is used to get the surrounding temperature and humidity in the greenhouse, and NodeMCU is used as the main microcontroller. Some other facilities such as fan and heater are used to adjust the inside environment. The system could monitor the growth environment continuously with Internet-connected, the monitoring data is transmitted and stored in the ThingSpeak cloud, the users can visualize the live data through a webpage or phone APP in real-time. If the environment condition is out of the predefined level, the environment is monitored continuously, and the system can be adjusted automatically. This system can be deployed in the greenhouse simply and maintain the greenhouse environment in a normal range dynamically and continuously.

**Keywords**—Smart greenhouse; ThingSpeak cloud; NodeMCU

## I. INTRODUCTION

As urbanization is continuously expanding, resulting in a huge decrease in arable land, the rapid growth of human population has increased the demand for food as well. The traditional agriculture method is not satisfied with current demands [1]. The problem of food for humans is becoming more and more serious. Food security is still a main issue in current days, it is an integrated and long-term task to deal with not only for agriculture but also for political will [2]. Climate change has resulted in severely damaging agroecosystems of the Loess Plateau in China, further aggravating the loss of soil and crop yields [3]. The living standards of people are greatly affected. The traditional farming model requires more land and manpower to manage, so traditional farming along could not be sufficient and resolve food security problems, it requires to apply of modern technology especially the Internet of Things (IoT) to improve it in this digital age.

In nowadays, the greenhouse is applied widely in the countryside to plant crops or vegetables for the whole year regardless of the seasons. It has heat-keeping, anti-coldness, and transparency characteristics. The main significance of greenhouse is the climate inside can be controlled at a suitable level constantly of the specific plant favorable. Some important nursery factors such as temperature, humidity, soil moisture,

pH, light intensity et al [4]. Along with the yield prediction character, the most efficient production of the greenhouse could be possible with the help of advanced technology. Researchers and Engineers use Internet of Things (IoT) and other modern technologies to make it realize. With the popularity of smartphones, farmers could use phone to monitor and control the greenhouse in real-time without extra human intervention [5]. IoT technology applied in agriculture is a developing trend, the potential benefits not only expand the yields and quality of the planting crops but also reduce farmers' burden and improve income.

It is important to apply smart greenhouse technology in urban areas. An Arduino uno based smart greenhouse prototype was designed and implemented, the greenhouse environment is monitored in real-time and can be accessed through an Android application. The user also could use the Android phone to manually control the inside environment remotely. The prototype was examined and highlighted that it could improve the yields of plants [6]. An STM32-based temperature monitoring and control system was developed; this development proved that smart agriculture could ease the management burden and increase the yields of crops [7]. Artificial intelligence (AI) technology is playing an important role in the smart greenhouse as well. An improved fuzzy neural network algorithm was designed to fit the intelligent greenhouse development. It is a trend that different technologies such as 5G, AI, NB-IoT, and Cloud should be applied to make the greenhouse more sustainable and smarter [8]. The greenhouse can be designed and implemented in a modern way by using different kinds of technologies.

The rest of this paper is divided into sections mentioned below: Section II is about the literature review, it summarizes the corresponding works of the smart greenhouse. In Section III, the proposed system and experimental setup are presented in detail. In Section IV, the results of this research project are described. Finally, the conclusion is summarized in Section V.

## II. LITERATURE REVIEW

Though IoT technology is widely used in different fields, such as smart parking systems, smart healthcare and so on, it still does not apply in large-scale agriculture in many countries especially developing countries. A Lora-based small-scale smart greenhouse was developed; the system could monitor soil moisture, light strength and temperature; the data was transmitted to the Tata server; and the data could be retrieved

from Microsoft Azure Cloud and displayed on the developed webpage through the network [9]. In order to further efficient management of farming, a camera was deployed not only to monitor the growth conditions of plants, but also to check which disease the crops have by using image processing technology [10]. An Esp8266-based smart and automated controlling agriculture system was designed, four parameters: temperature, humidity, light and soil moisture were monitored in real-time, the data was transferred to the customized webpage through a wireless network, the system also could adjust its environment conditions automatically if one factor was out of the predefined threshold so that it could maintain the optimum environment for the crops to grow rapidly [11]. A data analysis platform based on docker technology was designed and implemented; this platform was deployed simply regardless of the underlying operating system [12]. Data analysis is an important process after data gathering in IoT technology.

The yields of outside crops are mostly influenced by severe weather such as rain, and storm. The change of temperature and humidity could result in different diseases for the crops. An Arduino Nano-based smart agriculture system was developed to realize monitoring and controlling of the greenhouse in real time, in this study, compared with traditional monitored parameters: temperature, soil moisture and light intensity, one more rain sensor was used to detect the weather conditions, and to trigger the top of the greenhouse open or close automatically to irrigate the crops. Users not only could see the monitoring environment values on a 16X2 LCD display, but also access the data through the designed phone application remotely [13]. An intelligent supervisory fuzzy controller (ISFC) was developed to control and adjust the greenhouse environment remotely [14]. Soil is the base for plants to grow, so an intelligent soil management system is designed and developed, it is more convenient for farmers to manage and maintain their crops in the greenhouse [15]. Farmers could observe the monitoring data through a blynk application. In order to quickly identify the vegetable disease, deep learning algorithms were integrated into the smart greenhouse monitoring system to avoid loss in the early stage [16]. With the help of machine learning technology can make the whole smart greenhouse system more intelligent [17]. Solar power could be used as a renewable resource to provide electricity for the whole smart greenhouse monitoring system. It not only reduces the cost, but also no pollution to the environment. And the extra power was stored in the rechargeable battery to save cost [18]. A smart greenhouse system can be integrated with hydroponics planting, the soil is saved in this way, and it could provide rich nutrients to make the vegetables grow freely and rapidly [19]. With the help of advanced technology, the smart greenhouse is executable and efficient.

### III. PROPOSED SYSTEM

Internet of things (IoT)-based technology applications are a tendency to make everything intelligent and facilitate. In this project, a low-cost and sustainable smart greenhouse monitoring and controlling system for agriculture is designed and developed. The architecture and experimental prototype setup of this research project are presented in this section.

#### A. System Architecture

In this project, the greenhouse environment not only can be monitored in real-time, but also it can adjust the environment conditions: temperature and humidity at a suitable level automatically and continually. If the temperature is lower than the predefined threshold, the heater would turn on, if the temperature is higher than the preset temperature or the humidity is outside the threshold, the fan would turn on, otherwise, the fan and heater are turned off. Users can remotely access the monitoring data by phone or webpage through the network wherever the users are. The proposed architecture of the smart greenhouse system is illustrated in Fig. 1.

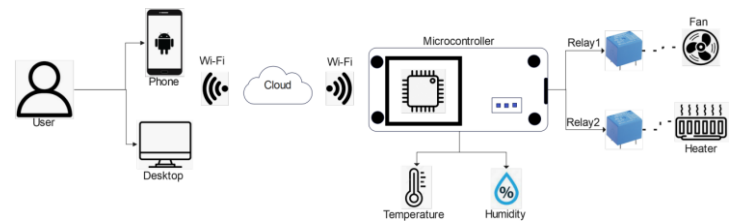


Fig. 1. System Architecture.

#### B. Experimental Setup

This project is developed based on a NodeMCU microcontroller coupled with other electric devices. NodeMCU is a low-cost and open-source Internet of Things developing platform. Arduino IDE is used as the programming tool; the firmware development is based on the ESP8266 environment. The NodeMCU diagram is shown in Fig. 2.



Fig. 2. NodeMCU Microcontroller.

HT22 sensor is used as the humidity and temperature detecting node. It has 4 pins: VCC, DATA, NC, and GND. The normal voltage range of VCC is from 3v to 5v. The working principle of DHT22 is through a built-in capacitive humidity sensor and thermistor to measure. The data is transmitted to the controller through the DATA pin. The maximum working current is 2.5mA [20]. The picture of the DHT22 sensor is shown as Fig. 3.





Fig. 3. DHT22 Sensor.

According to the proposed system architecture, a small-scale and easy-installed smart greenhouse monitoring and controlling system prototype is designed and implemented. In this system NodeMCU is applied as the main microcontroller to collect and transfer the monitoring data, a DHT22 sensor is used to get the temperature and humidity of the environment, the collected data is stored in the ThingSpeak cloud, the data can be visualized through a mobile application or ThingSpeak website through specific channel ID. A two-channel relay is used as the actuator to connect the heater and fan. The diagram of the relay is shown as Fig. 4.



Fig. 4. 2-Channel Relay.

A 16x2 LCD is used to display the current monitoring environment values. The connection of the hardware modules is shown as Fig. 5.

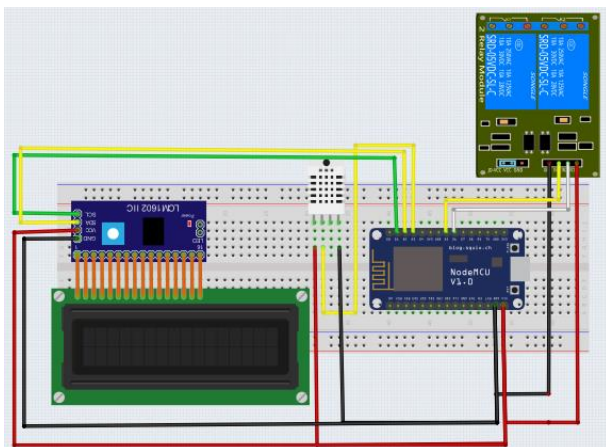


Fig. 5. Connection of Hardware Modules.

The prototype of the proposed system is designed and implemented, and is shown as Fig. 6. LCD display, DHT22 sensor, 2-channel relay, switch and NodeMCU microcontroller are deployed on the top of the printed circuit board (PCB) board. Two 18650 batteries are used to support power for this system and put on the back of the PCB board. The relays connect to the fan and heater. The LCD display screen is connected to the microcontroller via an I2C interface. The real-time monitoring data can not only be displayed on the LCD display screen but also can be visualized by phone APP and webpage through network.

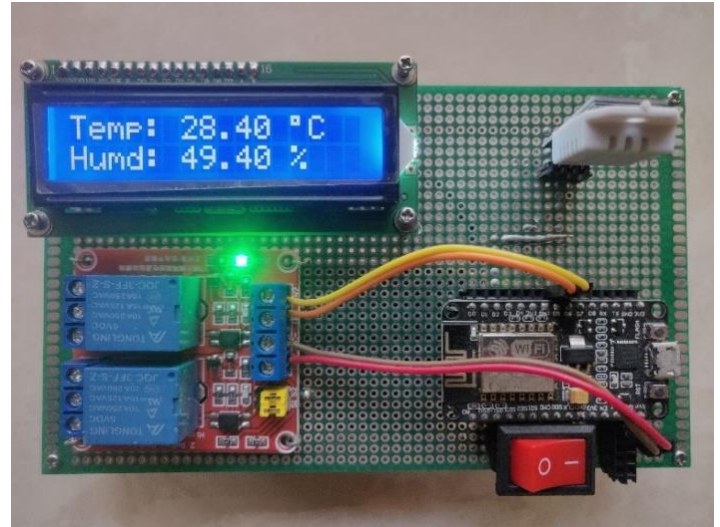


Fig. 6. Implementation of Proposed System.

#### IV. RESULTS

The data also can be assessed through a mobile application called ThingView or ThingSpeak website remotely with an Internet connection. The monitoring data from the ThingSpeak webpage is shown as Fig. 7. From the platform, users can monitor the greenhouse environment in real-time. The live data is accessed through the smartphone application with a specific channel ID and is shown as Fig. 8. The data is updated every 15 seconds. The greenhouse environment is adjusted at a predefined suitable level dynamically.

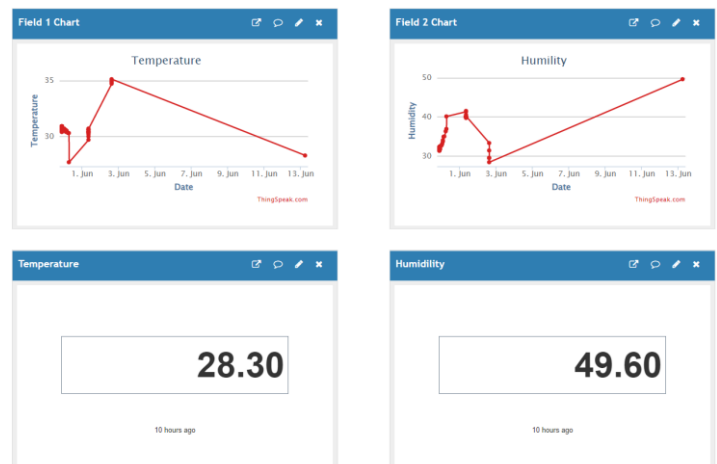


Fig. 7. Real-Time Monitoring Data through Webpage.



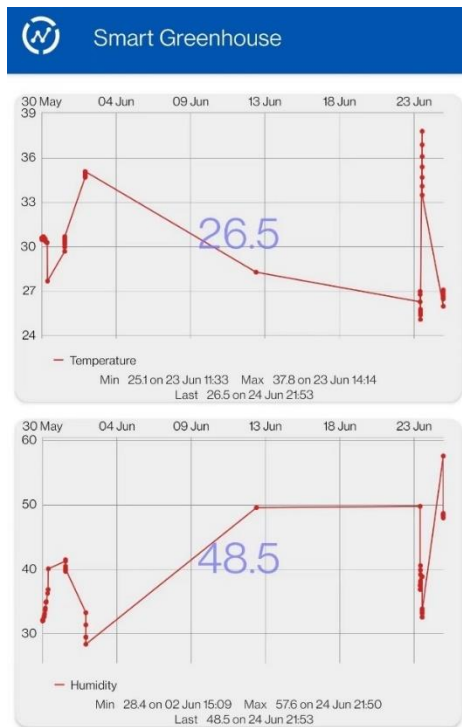


Fig. 8. Real-Time Monitoring Data through Phone APP.

## V. CONCLUSION

Based on the research question, data analysis is a useful way; from the simple test demonstration we can conclude that there is a negative relationship between temperature and humidity. In the future, machine learning technologies such as clustering and classification algorithms can be applied in this system, based on the big data and machine learning technology the expected model could be constructed accurately, and the model can decide whether to activate the specific instrument to predict and adjust current condition for the whole greenhouse system. It adjusts the environment in a normal range dynamically and automatically, to avoid big losses and reduce human intervention.

## ACKNOWLEDGMENT

The authors would like to thank to the support of Information Engineering School, Henan Vocational University of science and technology, for their efforts to conduct this research project. The authors would like to thank to Professor Dr. Omar Ismael Ibrahim for his valuable suggestions.

## REFERENCES

- [1] Perla M.F., Oscar A.J., Enrique R.G. et al., "Perspective for aquaponic systems: "Omic" technologies for microbial community analysis," *BioMed Research International*, vol. 2015.
- [2] Prosekov, Alexander Y., and Svetlana A. Ivanova. "Food security: The challenge of the present." *Geoforum* 91 pp.73-77, 2018.
- [3] Li, Zhi, Wen-Zhao Liu, Xun-Chang Zhang, and Fen-Li Zheng. "Assessing the site-specific impacts of climate change on hydrology, soil erosion and crop yields in the Loess Plateau of China." *Climatic Change* 105, no. 1 pp. 223-242, 2011.
- [4] Vimal, P. V., & Shivaprakasha, K. S. (2017). IOT based Greenhouse Environment Monitoring and controlling system using Arduino

- platform. *2017 International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT)*. <https://doi.org/10.1109/icicict1.2017.8342795>.
- [5] Rezvani, Sayed Moin-eddin, Hamid Zare Abyaneh, Redmond R. Shamshiri, Siva K. Balasundram, Volker Dworak, Mohsen Goodarzi, Muhammad Sultan, and Benjamin Mahns. "IoT-based sensor data fusion for determining optimality degrees of microclimate parameters in commercial greenhouse production of tomato." *Sensors* 20, no. 22 (2020): 6474.
- [6] Kirci, Y. C. P., Erdinc Ozturk, and Yavuz Celik. "Smart greenhouse and smart agriculture." In *Conference of Open Innovations Association, FRUCT*, no. 29, pp. 455-459. FRUCT Oy, 2021.
- [7] Lv, Junyuan, Weifeng Kong, Yongle Shi, and Han Bao. "Design of Intelligent Greenhouse Management and Control System." *International Core Journal of Engineering* 7, no. 1 (2021): 230-232.
- [8] Liu, Haibin, Shengyu Fang, and Xinqin Guo. "Research and Design of Intelligent Greenhouse Control System Based on AIoT Fusion Technology." In *IOP Conference Series: Earth and Environmental Science*, vol. 474, no. 3, p. 032036. IOP Publishing, 2020.
- [9] Reka, S. Sofana, Bharathi Kannamma Chezian, and Sanjana Sangamitra Chandra. "A novel approach of iot-based smart greenhouse farming system." In *Green buildings and sustainable engineering*, pp. 227-235. Springer, Singapore, 2019.
- [10] Ardiansah, I., Bafdal, N., Suryadi, E., & Bono, A. (2020). Greenhouse monitoring and automation using Arduino: A review on Precision Farming and internet of things (IOT). *International Journal on Advanced Science, Engineering and Information Technology*, 10(2), 703. <https://doi.org/10.18517/ijaseit.10.2.10249>.
- [11] Kulkarni, Manasi R., Neha N. Yadav, Sanket A. Kore-Mali, and Saurabh R. Prasad. "Greenhouse automation using IoT." *International Journal of Scientific Development and Research (IJS DR)* 5, no. 4 (2020): 239-242.
- [12] Hyun, W., Huh, M. Y., & Park, J. (2018). Implementation of docker-based Smart Greenhouse Data Analysis Platform. *2018 International Conference on Information and Communication Technology Convergence (ICTC)*. <https://doi.org/10.1109/ictc.2018.8539551>.
- [13] Naik, M. R. Greenhouse Environment Monitoring and controlling through IOT. *International Journal for Research in Applied Science and Engineering Technology*, 10(6), 2412–2417,2022.
- [14] Aghaseyeddabollah, M., Alaviyan, Y., & Yazdizadeh, A. (2021). IOT based Smart Greenhouse Design with an intelligent supervisory fuzzy optimized controller. *2021 7th International Conference on Web Research (ICWR)*. <https://doi.org/10.1109/icwr51868.2021.9443022>.
- [15] Lanitha, B., E. Poornima, R. Sudha, D. David, K. Kannan, R. Jegan, Vijayakumar Peroumal, R. Kirubagharan, and Meroda Tesfaye. "IoT Enabled Sustainable Automated Greenhouse Architecture with Machine Learning Module." *Journal of Nanomaterials* 2022 (2022).
- [16] Fatima, Neda, Salman Ahmad Siddiqui, and Anwar Ahmad. "IoT-based Smart Greenhouse with Disease Prediction using Deep Learning." *International Journal of Advanced Computer Science and Applications* 12, no. 7 (2021).
- [17] Jaiswal, Himanshu, Ram Singuluri, and S. Abraham Sampson. "IoT and machine learning based approach for fully automated greenhouse." In *2019 IEEE Bombay Section Signature Conference (IBSSC)*, pp. 1-6. IEEE, 2019.
- [18] Porselvi, T. "Automatic Control and Monitoring Of Greenhouse System Using Iot." *Turkish Journal of Computer and Mathematics Education (TURCOMAT)* 12, no. 11 (2021): 2708-2715.
- [19] Nurjannah, D. R., Supriadi, D., Sutiawan, A., & Kustiawan, I. (2020). Designing smart greenhouse systems using SCADA based on IOT. *IOP Conference Series: Materials Science and Engineering*, 850(1), 012002. <https://doi.org/10.1088/1757-899x/850/1/012002>.
- [20] Yajjie Liu and Dr. R. Annie Uthra, Bluetooth Based Smart Home Control and Air Monitoring System, *International Journal of Advanced Research in Engineering and Technology (IJARET)*, 11(5), 2020, pp. 264-274.

# Design of Accounting Information System in Data Processing: Case Study in Indonesia Company

Meiryani<sup>1</sup>

Accounting Department, School of Accounting, Bina Nusantara University, Jakarta, 11480 Indonesia

Gabrielle Beatrice Hidayat<sup>4</sup>

Accounting Department, School of Accounting, Bina Nusantara University, Jakarta, 11480 Indonesia

Jessica Paulina Sidauruk<sup>7</sup>

Accounting Department, School of Accounting, Bina Nusantara University, Jakarta, 11480 Indonesia

Dezie Leonarda Warganegara<sup>2</sup>

Management Department, BINUS Business School, Doctor of Research in Management, Bina Nusantara University, Jakarta 11480, Indonesia

Erna Bernadetta Sitanggang<sup>5</sup>

Accounting Department, School of Accounting, Bina Nusantara University, Jakarta, 11480 Indonesia

Mochammad Fahlevi<sup>8</sup>

Management Department, BINUS Online Learning, Bina Nusantara University, Jakarta 11480, Indonesia

Agustinus Winoto<sup>3</sup>

Accounting Department, School of Accounting, Bina Nusantara University, Jakarta, 11480 Indonesia

Ka Tiong<sup>6</sup>

Accounting Department, School of Accounting, Bina Nusantara University, Jakarta, 11480 Indonesia

Gredion Prajena<sup>9</sup>

Affiliation: Computer Science Department, School of Computer Science, Bina Nusantara University, Jakarta, Indonesia 11480

**Abstract**—This study aims to determine the implementation of System Application and Product in Data Processing (SAP) in a company to provide solutions for companies to obtain reliable reports and improve performance in a company. This study uses the mixed method through interviews with resource persons who have work in well-known company. The data obtained were analyzed by the method of literature study from data on the internet. The results of this study indicate that many companies still apply manual systems in reporting, one of them is the lack of adequate technological facilities within the company so that companies cannot fulfill their business processes optimally due to not using integrated system that connected with each other.

**Keywords**—Accounting; information systems; SAP; ERP; implementation of SAP

## I. INTRODUCTION

The world is currently growing in the world of technology is needed by companies to increase efficiency and achieve maximum accuracy. In the current era of digitalization, technological developments in the world are increasingly advanced, where these developments are increasingly modern and lead to the digital world, so that this has a major impact on various fields and sectors of activity [1]. Many companies especially in the business sector have grown and expanded, giving rise to a drastic increase in business competition. In its development, several companies demand to obtain accurate and relevant data and information in their business processes [2].

In the past, before the company technology used a manual recording system, the drawback of this manual system is that fraud and human error often occur, so that it is considered less effective and efficient. In addition, companies are often found that they cannot fulfill their business processes optimally [3],

so that their records still often occur. The problem that occurs is because there are still many companies that do not have a system that is not connected to each other to get relevant reports [4].

According to Rosenbom there are five strategies in increasing company productivity that are effective and efficient that can be applied [5], namely changing management rules, changing the nature and composition of inputs, increasing new technology, increasing new products, and expanding new markets. Information systems can support two of the five methods above, namely multiplying new technologies and expanding new markets. The solution is to apply System Application and Data Processing (SAP) software which combines the two systems, namely accounting and management information systems that can meet the needs of all parts of the company. Over the last few years, several companies have switched to using applications or software known as System Application and Data Processing (SAP) which are considered to be able to meet business needs more efficiently and effectively so that it is easier for all departments to run the company.

System Application and Data Processing (SAP) is an enterprise resource planning (ERP) software. ERP is an application that can integrate to meet the company's operational needs efficiently which consists of various sets of modules such as manufacturing, finance, HRD, material management, sales, and distribution that are connected into one database. Therefore, System Application and Data Processing (SAP) has developed and is recognized by manufacturing companies, because it is considered to greatly increase effectiveness and efficiency in various matters related to company operations because it is only integrated in one

software so that the settings will also be easier compared to using manual way [6]. This system is expected to provide solutions for companies to deal with all problems that occurred previously and can help employees to work more efficiently. In addition, SAP has other uses such as being able to improve corporate governance data to provide confidence to investors by looking at overall performance through a real time transaction system [7].

In recent years, it is being discussed that many companies have applied SAP to help meet their business needs. According to data from reference sites, there are many ERP systems emerging, but as many as 80% of companies in Indonesia have applied SAP as their ERP system to facilitate their business processes [8]. Several leading companies in Indonesia have applied ERP including PT Pertamina, Starbucks, Astra International, Erafone, Bank Mandiri, PT Garuda Indonesia, Telkomsel, Blue Bird and many others. Digital transformation is not just an alternative but a must for all lines of the company [9].

Three phenomena of applying SAP in the company: First, in about 50 years ago, the most phenomenal technological development was the emergence of the internet, where the internet is a technological development that combines telecommunications and computer technology [10]. Therefore, many companies are now applying SAP in their companies to be able to compete nationally and internationally. So that all companies are required to keep up with global developments by adopting more capable technology such as using SAP as a new breakthrough so that the company can maintain its existence and be able to compete at the national or global level [11]. Second, according to Billyan and Irawan the phenomenon of applying SAP has begun to spread throughout Indonesia, both from service and manufacturing companies, because the use of ERP itself can be used to analyze the consequences that hinder all processes within the company can be investigated [12]. Third, the phenomenon of the application of SAP has been applied to companies in Indonesia and outside Indonesia. This happens because there are some problems in receiving information so far it is considered less than optimal because it is considered less effective, one of which is the application of information technology that is less than optimal. In a company, the company needs information that can be used to make good and fast decisions. Therefore, through this background the researcher trying to find a solution for companies that combine data and technology to record systemically to get reliable reports.

Based on the description above, this research intends to study "Design of Accounting Application Information System and Product in Data Processing in Indonesia Company". Researchers want to examine whether the application of SAP in the company can help productivity and efficiency in the company's operational processes.

- a) Is SAP really needed by the company?
- b) Is SAP able to reduce fraud and human error and increase efficiency in companies that apply SAP
- c) Is the costs incurred to install SAP be balanced with the results obtained by the corporation in terms of accuracy.

## II. LITERATURE REVIEW

### A. Grand Theory

The researcher raised one of the theories that could underlie the formulation of the problem, namely by using the Agency Theory [13]. Agency Theory is a theory that explains the relationship between the two parties, namely the company management (agent) and company owner (principal). Under certain conditions, the owner of the company, namely the principal, always needs information related to the company's activity processes. Through reports that have been prepared by the agent, the company owner (principal) can also receive the information need and provide an assessment of performance within a certain time. In research, there is a discussion that is under agency theory, that statements regarding understand the problems that occur between company owners and agents in giving reports in a company [14].

### B. Accounting Information System

Accounting Information System is the root to obtain information quickly and reliable. Fast means that the information obtained is proven to be really actual and accurate the time. While accurate is based on adequate and reliable evidence accountable for the truth. The presence of an accounting information system can help company to obtain reliable information, and obtain information which is useful in making a good decision under certain conditions. The following describes the notion of an accounting information system based on the opinion of some experts, namely:

a) Bodnar and Hopwood [15], an accounting information system is a combination of resources, such as individuals and equipment, designed to convert financial and other data into useful information to various parties in making a decision.

b) Widjajanto [16], the accounting information system is a layer of various documents, communication tools, personnel implementers, and various reports designed to transform data financial report into financial information.

Based on the above definition, it can be concluded that the accounting information system is combination of resources such as individuals and equipment designed to manage financial data and other data into useful information in making decisions in controlling, planning, and managing the organization.

According to Azhar Susanto listed in his book consists of 6 (six) Accounting Information System indicators [17], namely:

a) Hardware is hardware that is used to combine, process, store, enter, and produce data processing things to provide information some information.

b) Software is a combination of various programs that are used to process data application on the computer.

c) Database is a system used for data collection or writing by using computer media to support information so that it is always available in real time.

d) Procedures are various activities that are carried out repeatedly with techniques use the same method. Consistency is the main key in the process of an organization.

e) Brainware is a human resource that participates in the process of making a product information system, which consists of combining, processing data, distributing data to the use of data for the needs of an organization or company.

f) Communication Network is the use of electronic means to transfer both information and data from one location to another.

### C. Enterprise Resource Planning (ERP)

According to Chofreh et al. [18], Enterprise Resource Planning (ERP) is an information system designed to integrate all company activities both internally and externally to the company that allows access to data quickly and reliably. This system includes manufacturing, distribution, personnel, project management, payroll, and finance. ERP is also a shortcut from technology information to assist companies in managing company operations by implement a shared database. So that the presence of ERP can support the productivity and operating efficiency of business processes by integrating business transaction management activities such as sales, manufacturing, marketing, finance, accounting, logistics, and resources. They also argue that there are three reasons why ERP systems are developing very rapidly lately, including the development of globalization, the era of the 2000s, and the need for information integration better.



Fig. 1. ERP.

The success of applying ERP in the company, according to Pabedinskaite [19] on the research he did in his book entitled "Factors of successful implementation of ERP systems". He revealed that there are 16 (sixteen) indicators in determine the success of ERP implementation, Fig. 1. Here are 16 factors in determining ERP implementation success:

- a) There is good communication from one department to another.
- b) Conduct optimal job training.
- c) There is support from the company's superiors.
- d) There is an equivalence between the company's business and the application of technology.
- e) There is employee involvement in the project.
- f) Reorganizing business processes.
- g) Good organization in transferring data.

- h) Competent external consultant.
- i) Organizational change management.
- j) Develop an organized and measurable plan.
- k) Conduct periodic and tight control on the implementation of ERP within the company.
- l) Involvement from management.
- m) Employee engagement.
- n) Measurable company targets.
- o) In accordance with the needs analysis of the organization.
- p) There is a good relationship with suppliers.

### D. System Application and Products in Data Processing (SAP)

According to Gunawan and Ikhsan [20], System Application and Products in Data Processing (SAP) is an ERP software that is integrated between its various modules such as SD (Sales Distribution), MM (Material Management), FICO (Financial and Controlling), HR (Human Resource Management), CO (Controlling), PP (Production Planning), and PM (Plant Maintenance) and others. Because the system supports integration, making software This is often used by many large companies in order to achieve their goals company, but by installing SAP in the company has a weakness, namely: requires a large amount of money, starting from taking care of the license, SAP training, software etc. SAP (System Application and Products in Data Processing) was first discovered in 1972 in Walldorf, Germany, by Dietmar Hopp, Hans-Werner Hector, Hasso Plattner, Klaus Tschira and Claus Wellenreuther who were former employees at IBM. SAP generally has three functions, namely: Functional (for background in finance, accounting, HR, ABAPer (for programmers), Basis (for admin work). The use of SAP is one of the most famous ERP software in the world and reliable by companies in the world including Indonesia because it has been recognized that SAP is the one of the most superior media and can be configured according to needs business. Here are the facts about SAP ERP:

Distribution of companies using SAP ERP by Country

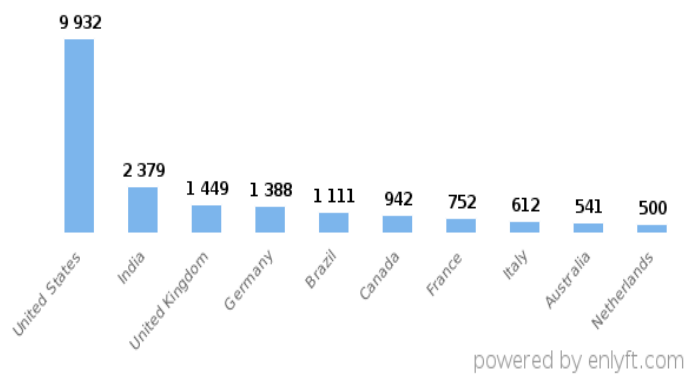


Fig. 2. Top Countries that used SAP ERP.

Based on the picture, Fig. 2, it shows that 36% of SAP ERP users are in the United States, 9% in India and 5% in the UK.

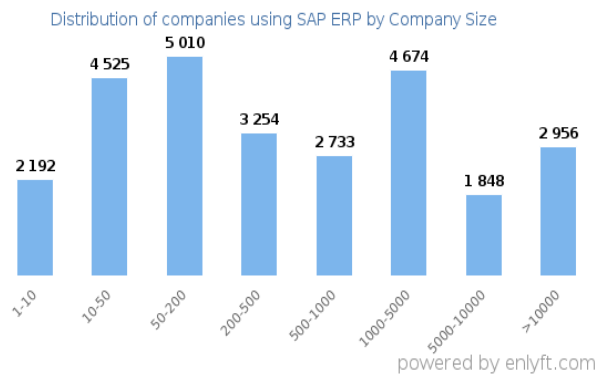


Fig. 3. Distribution of SAP ERP.

Based on the graph, Fig. 3, it shows that of all those who use SAP ERP, 25% are small customer (<50 Employees), 40% are middle, and 35% (>1000 employee).

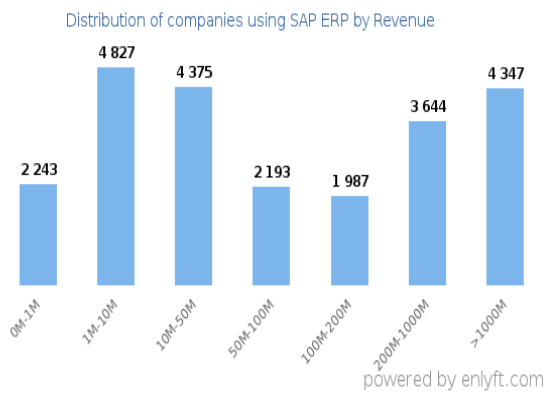


Fig. 4. Distribution of SAP (Revenue).

Based on the graph, Fig. 4 shows that customers who use SAP ERP, 47% are small customers (<\$50 Million), 10% are middle, and 33% are large (>\$1000 million).

#### E. Theoretical Framework

##### The Effect of SAP/ERP Implementation on the Company

SAP/ERP needs to be implemented by the company. As we know that when the world is now leading to the digitalization era, which means that everything has turned to technology sophisticated, so that it requires many companies to switch use technology to maintain its existence. In addition, it is often found that. Many companies still do not have a system that is connected to each other. So, it has weaknesses, one of which is the difficulty of connecting between one department to another, recording reporting is still common errors due to human error, inefficient, and so on. To that end, the application of SAP/ERP is very necessary to overcome problems in the business world because through. This platform is making a big impact for companies such as increasing productivity, efficiency, reliability, punctuality and so on. On the other hand, if the

company does not use existing technology such as SAP/ERP then the company will allow frequent errors and result in the company being unable to compete on a global level. According to Jacobs and Whybark [21], “ERP is the technology e-business, an enterprise-wide transaction framework with links to processing sales orders, inventory management and control, planning and finance production and distribution”. ERP is a direct result and extension of manufacturing resource planning and as such, covers all MRP II capabilities. ERP is superior because:

- a) Implement a set of resource planning tools across all companies.
- b) Presents real-time system integration of sales, operations, and financial data.
- c) Linking resource planning approaches to extended supply chains customers and suppliers.

### III. RESEARCH METHODOLOGY

According to Saunders et al. [22], the object of research is an attribute that explains about what and who is the object, where, and when the research is carried out. According to Sekaran and Bougie [23], the object of research is the names of research variables, refers to the identification of problems, hypotheses, and definitions contained in the previous chapter. The object of research here includes the SAP software. Method is derived from the Greek, namely *methodos*, which means way. Besides it's a method that comes from the Greek, *metha* (to pass or through), which means the way or way that must be traversed to achieve a goal. Whereas research which examines the problem by collecting facts. Method that researchers used for this research is qualitative method. Qualitative method is research that aims to understand the phenomena experienced by research subjects for examples of behavior, perceptions, motivations, actions and including in the type of qualitative method that uses data results in the form of sentences that can answer the problem formulations such as “what”, “how”, and “why”.

This type of research is casual research. Casual research aims to knowing the causal relationship that occurs from each variable to get facts from a phenomenon and seek factual information about the application of the SAP. According to Lind, Marchal, and Wathen [24], the general environment in the organizational environment is a broad external condition that can affect the organization and affect indirectly to organizational performance. According to Nurunnabi [25], the macro external environment includes several factors including economic, political and legal conditions, socio-culture, demography, technology and global conditions that affect the organization. General environmental changes may not have a major impact on environmental change; however managers still have to pay attention when planning, organizing, directing and controlling business activities:





Fig. 5. The Star Model.

In the Star Model, Fig. 5, it is shown that there are several variables related to the implementation of SAP. These include Human Resources Management related to the Reward System and then connected to the Business Processes and Structures generating Strategy. The components in SAP include the following:

- 1) SAP Financial Accounting
- 2) Controlling (CO)
- 3) Human Capital Management (HCM)
- 4) Production Planning (PP)
- 5) Project Systems (PS)
- 6) Sales and Distribution (SD)
- 7) Materials and Management (MM)
- 8) Quality Management (QM)
- 9) Plant Maintenance (PM)

#### IV. DISCUSSION

The author conducted module research in SAP to see if this system has been appropriate in the organization. The example below is one of the modules in SAP Hana where there are CO and FI modules in the finance module. Master data must pre-set according to company needs. After the master data has been completed is set, the accounting party does a business mapping and the output is universal journal, Fig. 6.



Fig. 6. Central Finance System Landscape.

We can see more about the FI posts as shown below that the activities manufacturing can be recorded in a system to be processed into information. For example, when the warehouse receives physical inventory, this must be recorded properly in SAP. GR (MIGO) must be made when the goods have been received in good condition. This has an effect when the invoice arrives, the accounting will call the PO number to verify payment. In the initial setup stage of the SAP system, we must ensure that the cost element, cost center, orders, and profit center are correctly summarized in SAP. Consistency is required in this is because we need to ensure that historical data can be used as a basis for do the analysis.

Likewise with the initial settings on the manufacturing side, we need to ensure that the physical stock and stock in SAP is correct and suitable. From the accounting side, we need ensure that the initial settings for tax calculations are in accordance with tax regulations that apply in Indonesia. It's better if the initial SAP settings are conditioned to be able to load supporting documents systemically. This will avoid a long time to search hardcopy documents.

The next stage is to determine the user in each SAP module. This is important because Segregation of Duties needs to be prioritized, meaning that everyone has a responsibility Answer each according to their job role. Therefore, access to SAP too tailored to the needs of each user. Not everyone is granted access as "superusers". A review of the Segregation of Duties should be carried out once a year. This is to avoid fraud. The company's advantages in using SAP as an ERP are as follows:

- a) There are various types of modules in SAP that have the ability to: supports all transactions and each of these modules will work in conjunction with one another with the others.
- b) SAP is also supported by a NetWeaver platform that supports development and logistics software.
- c) SAP has a programmed, which will make it easier for developers to implementation of business logic.
- d) All data in SAP can be stored in 1 server for a long period of time and can be accessed by various parties when needed.
- e) SAP can be modified according to company needs. This means that the modules in SAP are not a locked module but a module that can be modified.
- f) Allows integration globally.
- g) Reducing the level of complexity of applications and technology.
- h) Helping the smooth supply chain and integrating the results into one with financial reports.
- i) Reducing the need to update data continuously.
- j) Facilitating communication relationships internally and externally, inside and outside company.

In field practice, we can find that there are still many users who do not understand how to use the SPA. Usually, SAP will provide a session training and training so that users can be more familiar with using SAP. The following preparations need to be done before going live:



- a) Project Preparation, team formed for SAP project initial planning.
- b) Business Blueprint, establish a shared understanding of how the company intends to implement SAP in support of their business.
- c) Realization, in this phase, the implementation of the standard SAP methodology is carried out in two ways: packages base configuration (main scope) and final configuration (remaining scope). During this phase the solution is also tested.
- d) Final Preparation, the goal is to complete the final preparations including technical testing, final user training, system management and migration activities.
- e) Go Live Support, this phase is moving from the pre-production environment to the production phase.

However, the SAP system also has drawbacks, including:

- a) The implementation time and costs are not small.
- b) If careful preparation is not made, then the implementation process will become delayed.
- c) Users are not necessarily prepared to receive and operate SAP in time short

## V. CONCLUSION

Based on the description above regarding the application of SAP in the company, show that SAP can optimize business processes so that it can obtain information with fast and easy because it only has one integration that connects one with the others. SAP can help companies to reduce fraud and errors, so employee productivity increases. SAP will have a positive and significant impact on decision making. Looking at the current state of development, it shows that the world is getting more advanced and using very advanced technology. For this reason, the application of SAP in the company is a reliable solution for companies because it is only integrated in one system. SAP presence also provides many benefits for businesses, such as being able to help overcome unresolved business problems, get information in real time, reduce fraud or errors, and can obtain information quickly.

## REFERENCES

- [1] C. Leong, F. T. C. Tan, B. Tan, and F. Faisal, "The emancipatory potential of digital entrepreneurship: A study of financial technology-driven inclusive growth," *Information & Management*, vol. 59, no. 3, p. 103384, Apr. 2022, doi: 10.1016/J.IM.2020.103384.
- [2] S. S. Zhang, R. Riordan, and C. Weinhardt, "Interactive data: technology and cost of capital," in *Accounting Information Systems for Decision Making*, Springer, 2013, pp. 233–247.
- [3] Meiryani, "The influence of business process and management support on accounting information system," *Journal of Engineering and Applied Sciences*, vol. 12, no. 23, pp. 7416–7421, 2017, [Online]. Available: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85048341058&partnerID=40&md5=551f6caa54ecb51409612e1e2a3faaf9>
- [4] N. Juhandi, S. Zuhri, M. Fahlevi, R. Noviantoro, M. Nur Abdi, and Setiadi, "Information Technology and Corporate Governance in Fraud Prevention," in *5th International Conference on Energy, Environmental and Information System, ICENIS 2020*, 2020, vol. 202. doi: 10.1051/e3sconf/202020216003.
- [5] D. Rosenbaum, *Effectiveness, equity, and efficiency in community policing*. Sage Publications, 1994.
- [6] A. Purwanto et al., "Lean six sigma model for pharmacy manufacturing: Yesterday, today and tomorrow," *Systematic Reviews in Pharmacy*, vol. 11, no. 8, pp. 304–313, 2020, doi: 10.31838/srp.2020.8.47.
- [7] S. S. Halbouni, N. Obeid, and A. Garbou, "Corporate governance and information technology in fraud prevention and detection," *Managerial Auditing Journal*, vol. 31, no. 6/7, pp. 589–628, 2016.
- [8] R. Leonardo and T. A. Napitupulu, "Analysis of the Successful Implementation of SAP Business One in PT. PR Indonesia," *Budapest International Research and Critics Institute (BIRCI-Journal): Humanities and Social Sciences*, vol. 5, no. 3, 2022.
- [9] S. E. Sri Adiningsih, *Transformasi ekonomi berbasis digital di Indonesia: lahirnya tren baru teknologi, bisnis, ekonomi, dan kebijakan di Indonesia*. Jakarta: Gramedia Pustaka Utama, 2019.
- [10] F. R. Jacobs and D. C. Whybark, *Why ERP?: A primer on SAP implementation*, vol. 31. Irwin/McGraw-Hill New York, 2000.
- [11] B. Syaiful and W. Gunawan, "Assessing Leading ERP-SAP Implementation in Leading Firms in Indonesia," in *Journal of Physics: Conference Series*, 2017, vol. 801, no. 1, p. 012032.
- [12] B. F. Billyan and M. I. Irawan, "Analysis of Technology Acceptance of Enterprise Resource Planning (ERP) System in The Regional Office of PT. XYZ Throughout Indonesia," in *Journal of Physics: Conference Series*, 2021, vol. 1844, no. 1, p. 012008.
- [13] D. A. Bosse and R. A. Phillips, "Agency Theory and Bounded Self-Interest," *Academy of Management Review*, vol. 41, no. 2, pp. 276–297, Dec. 2014, doi: 10.5465/amr.2013.0420.
- [14] M. C. Jensen and W. H. Meckling, "Theory of the firm: Managerial behavior, agency costs and ownership structure," *J financ econ*, vol. 3, no. 4, pp. 305–360, 1976, doi: 10.1016/0304-405X(76)90026-X.
- [15] G. H. Bodnar and W. S. Hopwood, "Sistem informasi akuntansi," Jakarta: Salemba Empat, 2006.
- [16] N. Widjajanto, *Sistem informasi akuntansi*. Jakarta: Erlangga, 2001.
- [17] A. Susanto, "Sistem Akuntansi Prosedur dan Metode," BPFE, Yogyakarta, 2009.
- [18] A. G. Chofreh, F. A. Goni, J. J. Klemes, M. N. Malik, and H. H. Khan, "Development of guidelines for the implementation of sustainable enterprise resource planning systems," *J Clean Prod*, vol. 244, p. 118655, 2020.
- [19] A. Pabedinskaitė, "Factors of successful implementation of ERP systems," *Ekonomika ir vadyba*, no. 15, pp. 691–697, 2010.
- [20] W. Gunawan and R. B. Ikhsan, "Assessing ERP SAP implementation in the small and medium enterprises (SMEs) in Indonesia," in *Journal of Physics: Conference Series*, 2018, vol. 978, no. 1, p. 012013.
- [21] F. R. Jacobs and D. C. Whybark, *Why ERP?: A primer on SAP implementation*, vol. 31. Irwin/McGraw-Hill New York, 2000.
- [22] M. Saunders, P. Lewis, and A. Thornhill, *Research Methods for Business Students*, 5th ed. London: Prentice Hall, 2009.
- [23] U. Sekaran and R. Bougie, *Research methods for business: A skill building approach*. New York: John Wiley & Sons, 2016.
- [24] D. A. Lind, W. G. Marchal, and S. A. Wathen, *Statistical Techniques in Business & Economics*, 17th ed. New York: McGraw Hill Education, 2018.
- [25] M. Nurunnabi, "The impact of cultural factors on the implementation of global accounting standards (IFRS) in a developing country," *Advances in Accounting*, vol. 31, no. 1, pp. 136–149, 2015.

# MOOC Dropout Prediction using FIAR-ANN Model based on Learner Behavioral Features

S. Nithya<sup>1\*</sup>

Research Scholar, Department of Computer Science  
SRM Institute of Science and Technology  
Ramapuram Campus, Chennai, India

Dr. S.Umarani<sup>2</sup>

Professor, Department of Computer Science  
SRM Institute of Science and Technology  
Ramapuram Campus, Chennai, India

**Abstract**—Massive Open Online Courses (MOOCs) are a transformative technology in digital learning that incorporates new techniques through video sessions, exams, activities, and conversations. Everyone leads a successful life in their professional and personal skills learning courses during COVID-19. The research concentrated on employing video interaction analysis to characterize crucial MOOC tasks, including predicting dropouts and student achievement. Our work consists of merely generating and picking the best characteristics based on the learner behavior for evaluating the dropout measure. To locate the frequent objects for feature creation, an association rule-FP growth approach is applied. The neural network is implemented using frequent itemset-3, which is used for feature selection. The evaluation metrics are calculated by using the Multilayer Perceptron (MLP) method. The metric values were then compared to the proposed model and some base supervised machine learning models namely Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), K-Nearest Neighbor (KNN), and Naive Bayes (NB). The FIAR (Feature Importance Association Rule)-ANN(Artificial Neural Network) dropout prediction model was tested on the KDD Cup 2015 dataset and it had a high accuracy of over 92.42, which is approximately 18% better than the MLP-NN model. With the optimized parameters, we are solely focused on lowering dropout rates and increasing learner retention.

**Keywords**—Dropout prediction; data analytics; association rule mining; machine learning; artificial neural network

## I. INTRODUCTION

Massive Open Online Courses are the result of the blending of learning and the online world. It broadens the educational system by providing knowledge via modern internet technologies [1]. Moreover, an innovative instructional environment would assist learners in saving their learning time.

Billions of individuals utilize MOOCs for a variety of purposes, including professional growth, career transition, syllabus formation, secondary education, skills training, [2] and more. Courses are created by institutions for MOOC providers such as Coursera, edX, Udacity, and Swayam. Coursera, on the other hand, has done a far better job of preserving the pandemic boost than its closest MOOC competitors. From 31% in 2020 to 39% in 2021, Coursera's proportion of non-university courses expanded.

Learner enrollment in online classes has increased during the last decade. Simultaneously, learner dropout rates should

be raised. It's essential to consider the elements that contribute to inadequate retention rates [3]. The elements have to do with a learner's logs, access to materials, activities, and willingness to execute required actions in order to complete the program. Along with low dropout rates, course completion rates will rise.

This work has become focused on the analysis of learners' behaviors. In effective learning behavior, while the learner watches the video within the duration and submits the assessment on time. Moreover, analyzing the behavior of learners [4] with specific features such as accessing materials, discussion, course forum, and so on is significant.

Analytics has been popularized in recent years. The analytical approaches that extract relevant and meaningful enormous volumes of data and apply them to the educational system have revolutionized research [5] and it indicates as "Education Data Mining" (EDM), "Academic Analytics" (AA), and "Learning Analytics" (LA).

EDM has proven to be a good resource for revealing hidden information [6] and paradigms in data sources. Most educational institutions are being carried on e-learning platforms [7]. The huge amount of educational information recorded has laid the foundation for new research and analysis to better understand and increase learning performance.

Learning analytics is defined as the use of intellectual facts, learner-generated data, and models developed [8] to find knowledge and community interactions for prediction. To enhance learning progress and outcomes, learning analytics is associated with the gathering, processing, and interpretation of data sources. Learning Analytics predictive models employ any data mining, machine learning, or artificial intelligence technique, including classification, regression, prediction, and others, to evaluate the skills of learners, instructors, and universities.

Even though all attributes may be valuable in some circumstances, just a preferred number of attributes [9] are frequently used for identifying targets. In the KDD CUP 2015 dataset, an activity log is often used to collect several sorts of information regarding learners' behavior. In our research, behavioral variables were utilized to value the regularity of various learner behaviors, and we obtained the learners' parameters and used a variety of machine learning and

\*Corresponding Author.

artificial neural network classifiers to produce a significantly high prediction rate.

Establishing feature extraction and selection methods[10], as well as applying an artificial neural network to develop the model in learning analytics. The following are some key aspects of the proposed work:

For effective learning analytics, we proposed the FIAR-ANN model in our research. When association rule mining is utilized for feature selection, the efficiency of the process can be increased.

The relevance of the features is measured using an FP-Growth technique, which then brings them back to the highest values in the features.

The results of our FIAR-ANN model's examination on the KDD CUP 2015 datasets indicate that it is effective. Furthermore, we compare different metrics to the baseline models. Experiments were carried out to verify the research approach.

The layout of successful work in regards discusses past research that focuses on learner behavior to alleviate the problem of learner dropout and details the many aspects of the KDD CUP 2015 dataset in the third part, which includes our processes for cleaning and transforming the raw data to make it suitable for the analysis. In the subsequent sections, we evaluate the gathered data in terms of forecasting learner dropout and compare it to some baseline approaches. Finally, summarize the key achievements and suggest some simple directions for future research.

## II. LITERATURE REVIEW

Learners' performance prediction, behavior modeling, conversations, and retention have all been explored using data mining applications in online learning systems. Depending on the research setting, various studies have investigated multiple feature engineering strategies to extract additional sets of features.

The method of identifying a subset of relevant or essential features from raw data is known as feature selection, whereas feature extraction is the process of constructing a new variable from a collection of raw data.

On the KDD Cup 2015 dataset, Yafeng Zheng et al. [11] used the FWTS-CNN (Feature Weighting Time Series—Convolutional Neural Network) dropout prediction model, which used a decision tree to extract attributes out of a learner's record with a time series matrix [12] and then built a model from the weighted features using a convolutional neural network, and it had a better accuracy of over 87 percent.

Jing Chen and colleagues [13] developed DT-ELM (Decision Tree-Enhanced Learning Machine), a novel hybrid method that combines decision trees with extreme learning machines (ELM) that does not depend upon recurrent training. The first module, in particular, creates and extracts a number of attributes from learners' learning behavior records. The decision tree chooses features that can be classified well. It also gives the specified features more weight in order to

improve their categorization abilities. MATLAB R2016b and Python 2 are used to carry out the tests. The success of DT-ELM is proved by experimental results on the benchmark KDD2015 dataset, which show that it outperforms various basic ML models on several metrics by a percentile.

Cong Jin et al.'s[14] work starts with a feature extraction approach based on the students' behavioral content. Based on the Support Vector Regression parameters, an improved quantum particle swarm optimization (IQPSO) technique [15] is used to estimate the SDP (student dropout prediction) model. MATLAB 7.0 was charged with supporting the analysis with a 2:1 ratio of training and test subgroups at 10-fold cross-validation. The suggested SDP model outperforms benchmark models such as logistic regression (LR), back propagation (BP), and others, according to experimental results using public data.

As compared to earlier feature selection approaches, Anwar UIHaq et al. [16] designed a unique approach to anticipate greater results utilizing similarity multi-filter feature selection (MFFS). The feature ranking module discovers relevant characteristics, while the clustering module minimizes redundant features. Empirical results from a range of real-time data sources support the hypothesis that merging a feature picked using diverse distributed approaches leads to more resilient extracted features and improves the accuracy rate.

Bo Wei and colleagues [17] suggested a new optimization method assigned to particle swarm optimization with learning memory (PSO-LM). The learning recall policy's goal would be to get significant insights into those who are fitter and develop faster, although the genetic operation is frequently employed to reconcile on a small and large scale. By using the Weka tool to test the model's efficacy, each attribute must be scaled between 0 and 1, and the quality of each component must be assessed using the k-nearest neighbor classifier with 10-fold cross-validation. When compared to wrapper-based feature selection algorithms based on global standard datasets, the analysis revealed that they were more efficient.

Anupam Khan et al. [18] state that the most familiar educational data mining technique for determining pedagogical components of learning and assessing student achievement is association rule mining.

The classification algorithm may be used rigorously to establish a forecast connection, and the interactivity allows learners to correlate behavioral characteristics of their actions to program accomplishments.

In the field of educational data mining, Shaveen Singh et al. [19] explores the use of feature selection approaches combined with association rule mining to identify essential course activities and locate more notable links within these parameters. The task at hand is to come up with the right combination of learning activities that use various methodologies to achieve the course's intended learning results. Subsection formation, subgroup analysis, terminating condition, and outcome checking are the four basic phases included in it. The Communication and Information Literacy dataset UU100 enrolled 2,172 students and included a variety

of online actions. There had been no null data in the preprocessed dataset, which had 2172 occurrences with 19 features. To find valuable patterns and correlations among the features, an association rule mining technique is applied. The WEKA tool has been used to do data mining tasks by analyzing acquired data with various algorithms such as Naive Bayes, C4.5, and RBF Network, all of which have high prediction accuracy.

Abeer et al.[20]present a blended strategy for decreasing the high-dimensionality of DNA methylation data and extraction via the Kernel Density Estimation method, resulting in a considerably more accurate and quick-calculating method. The usefulness of the given hybridization technique is evaluated by the metrics of the proposed classifiers such as Naive Bayes, Random Forest, and SVM.

### III. MATERIALS AND METHOD

This section delves into the framework of the proposed model for predicting online learning dropouts. Feature extraction and selection techniques are used alone or in combination [21] to improve performance, such as projected accuracy, visualization, and concision of learned content. The benefit of feature selection is that crucial information about a particular feature is preserved. However, only a small number of qualities are needed, and the distinctive features are diverse. Some of them attempt to forecast a learner's performance in binary classes such as dropout or continue the Programme and so on.

#### A. Problem Definition

The learner dropout rate is the largest myth in MOOC development. Our work identifies the best attributes for predicting dropouts based on a learner's activity and uses a multilayer perceptron to measure accuracy. The available dataset for the KDD CUP 2015 originates from "XuetangX," China's largest MOOC platform[22] and a popular tool for predicting MOOC attrition. As displayed in Table I, the collection contains four catalogs from 79,186 students enrolled in 39 courses.

TABLE I. DATA ENACTMENT

Catalog	Depiction
Date	Timespan of each course
Object	Module in a course
log_train	Behavioral record
true_train	Insight into the actual data of the training set's enrollments.

#### B. Dataset Revelation and Preprocess

The KDD CUP 2015 public dataset, which contains 72,142 records with information about seven learning behaviors, as shown in Table II, including observing visual aids, retrieving objects, learner interaction, laying the course, closing the pages, analytical thinking, and surfing the web, is the exploratory data file used in this article. The binary classifier is used in the outcome analysis, with "0" indicating that students will complete their studies and "1" indicating that they will drop out.

#### C. Framework

The paper provides a strategy for dropout prediction using a FIAR-ANN model that integrates feature importance and neural networks. As illustrated in Fig. 1, the general stages of this approach are database cleaning, character separation, parameter estimation, and comparison.

The analysis begins with preprocessing the public dataset, then using association rules for feature selection to generate a frequent itemset based on behavioral attributes. The ultimate estimated values are then obtained using multilayer perceptron in the ANN model, and it was assessed using a variety of evaluation criteria.

1) *FIAR-ANN approach*: Fig. 2 displays the technical aspects of the FIAR-ANN model design, which are divided into two sections in this article: identifying and grading the characteristics in the activity log; deploying the ANN model; and evaluating the performance metrics.

#### D. Identification and Grading of Features

Our research features were derived from learners' learning activity logs. In an online-learning scenario, association rules can be useful since they can find correlations among distinctive characteristics in a dataset. It is being used to link learners' actions to their results in order to figure out what is influencing their learning chances favorably or adversely.

#### E. Association Rules

The goal of association rule learning is to establish significant relationships between items in huge datasets [23] using a rule-based machine learning system. Let  $Ar = (ar_1, ar_2...ar_n)$  denote a group of n elements, and  $Tr = (tr_1, tr_2...tr_m)$  denote a repository of m transactions.

Each transaction  $tr_i$  contains a subset of Ar's available items. A rule is defined as  $X = Y$ , where X, Y I, i.e., X and Y (also known as itemsets) are subsets of the accessible items. The precedent and subsequent rules are typically referred to as X and Y, respectively.

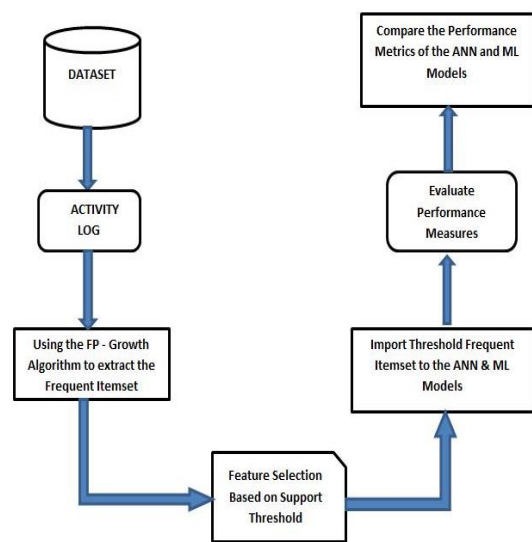


Fig. 1. The Layout of the Proposed Model.

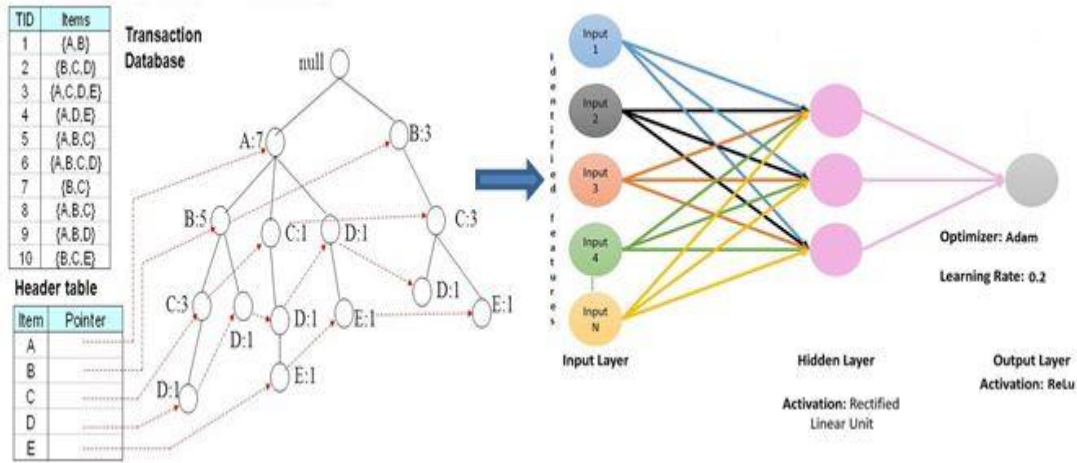


Fig. 2. Architectural Design of FIAR – ANN.

### F. Frequent Pattern (FP)-Growth Algorithm

The frequent pattern growth algorithm is an optimized and reliable substitute for the Apriori algorithm, which additionally searches the transaction database for frequent items.

To produce the association rules, this method uses a divide-and-conquer strategy[23]. The approach instead focuses on a data model known as the FP-tree that saves metadata on items and interactions. The transaction database is examined once to generate the FP-tree[21], and the group of candidate itemset F is then calculated and organized in support values.

Numerous measurements were presented to identify the significance and instructiveness of an association rule.

The sections that follow give two measures.

Support:

The support for a set of transactions Tr by an itemset is determined as follows:

$$\text{Support}(X) = |\{tr \in Tr; X \subseteq tr\}| / |Tr| \quad (1)$$

Confidence:

The proportion of the transaction's maximum set of frequent items and X values.

$$\text{Conf}(X \Rightarrow Y) = \text{Support}(X \cup Y) / \text{Support}(X) \quad (2)$$

### G. ANN Model

An Artificial Neural Network is made up of numerous neuron nodes [24] that are divided into three levels: input, output, and hidden layers. In current technological research and development, ANN approaches are most commonly used to identify feature sets that enable the revelation of suitable predictions [25], considering both the maximum of commonly used formal metrics and the understandability of the model's behavior for knowledge extraction from data collection.

We trained on 80% of the feature data set and tested on 20%. The proposed approach is defined as a multilayer-perceptron with single hidden layers, as seen in Fig. 2. The Rectified Linear Unit function is used to activate the neurons in the hidden layer. The input layer contains far more neurons than there are sources in the data set. However, the output layer contains only one neuron with a ReLu activation function, which is suitable for classification problems because it distributes actual content in the range of 0 to 1.

$$\text{relu}(z) = \max(0, z) \quad (3)$$

The computation costs have been reduced using an optimization strategy. With 10 epochs and a batch size of 10, the Adam optimizer developed to train artificial neural networks was employed and has reached the highest accuracy. Researchers really intended to broaden the outcomes of the research and establish that our method is applicable to future programs that emerge.

## IV. RESULTS AND DISCUSSION

Our work will look at the key elements of the FIAR-ANN model in this section, starting with creating frequent items using association rules and then inputting the selected parameters into the ANN model. The implementation of the FIAR-ANN model is broken down into two sections in this article: feature extraction and the enhanced ANN model. Fig. 2 depicts the process of implementation.

### A. Experimental Framework

#### 1) Software

- Windows 10 is the most popular version of Microsoft's operating system (Intel Core-i3 processor, 64-bit operating system).
- The Google Collaborator is a toolbox for the Scikit-learn Python package[26], and it was used to run the samples on a computer system with seven different sorts of events coming from two separate sources, as indicated in Table II.

TABLE II. DATA SOURCE ATTRIBUTES ARE USED IN THE ANALYSIS

Attributes	Proclamation
enrollment_id	A registered participant's unique identifier.
source	Actions based on the server and browser.
event	We devised seven distinct activities as
	Pbm - Putting intellectual tasks to use
	Video - Participants are observant of the filmed content.
	Access - Getting access to a lot of the learning resources.
	Wiki - Browse the internet for information.
	Discussion - Using discussion boards to exchange intelligence
	Navigate - Exploring several aspects of the software.
Page_close – Depart from the webpage.	
Output	The success (value 0) or failure (value 1) of a participant in a course

2) *Baseline models*: Our work applied five classic machine learning models as baseline models, especially Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), K-Nearest Neighbor (KNN), and Naive Bayes (NB), to offer a point of comparison for the outcomes of the FIAR-ANN model.

3) *Logistic regression*: The extended linear regression model is the basis of the supervised machine learning classification strategy used in logistic regression. It uses the regression coefficients of one or more components to calculate the likelihood of occurrence.

A strategy for estimating class-based predictor factors (x) is logistic regression [27]. It enables us to determine the possibility (p) of components of a specific class. A binary result is classified as a dropout or non-dropout in this work and it is represented as a 0 or 1.

The classic logistic regression model for evaluating the result of an occurrence, given a variable (x), is  $p = 1 / [1 + \exp(-y)]$ .

$$\text{In which } y = b_0 + b_1 * x \tag{4}$$

The exponential function is  $\exp()$ .

Given x, p is the probability of an event occurring.

The logistic function for the multiple parameters is as follows:  $b_0 + b_1 * x_1 + b_2 * x_2 + \dots + b_n * x_n = \log [p / (1-p)]$ .

The predictive beta coefficients are b0 and b1. Increases in x will be proportional to increases in p if b1 is positive. On the other hand, a negative b1 indicates that increasing x could result in a considerable decrease in p.

4) *Decision tree*: A decision tree leads to increased demand for developing and depicting forecasting tactics. It's

easy to learn and implement, and it's widely used for predictive analysis [28]. The basic goal is to divide a massive volume of information into smaller chunks. In predictive analytics, decision trees exhibit the prominent features in the datasets. The tree's root is at the top, with limbs descending. A node is a point on the limb wherever researchers divide the big bunch into shrinking units at every instance. A "leaf" is the term for the end node. In a decision tree, each limb represents a section, and each leaf node reflects the highlight attribute's result within a set range. The decision procedure begins at the root node, checks the related attribute of the item to be sorted, and chooses the outcome depending upon the level till it hits the leaf node.

5) *Random forest*: Bagging entails the use of many samples instead of a single sample. A trained model is a set of events that can be used to produce predictions. The decision trees' varied results make up the random forest algorithm. The final product will be chosen using a majority-voting procedure. Anomalies, as well as distortion, are less noticeable in Random Forest [29]. The Gini impurity is used for Random Forest class labels to minimize overfitting and bias errors, as well as prediction errors.

6) *KNN*: In the dropout prediction of this work, we utilize learner interaction within the dataset. In our KNN method, first choose a value for K. Using the Euclidean distance; calculate the distance between k neighbors [30]. Examine all of our neighbors to find which one is closest to our position. Our attribute is assigned to the class with the highest number. KNN looks for correlations between predictors and values within the dataset.

7) *Naive bayes*: According to the Naive Bayes classifier [31], the availability of one variable in a class appears to be unconcerned with the appearance of other variables in the same class [32]. It's simple to set up and especially handy for large amounts of data.

The Bayes theorem allows us to derive the posterior probability  $P(c|x)$  from  $P(c)$ ,  $P(x)$ , and  $P(x|c)$ .

Have a note of the following equation.

$P(c|x)$  denotes the posterior probability of a given class (c, target-dropout) given indicators (x, variables).

$P(c)$  is the class prior probability.

$P(x|c)$  denotes the probability of an indicator given a class.

$P(x)$  is the indicators prior probability.

8) *FIAR-ANN-Hyper parameters*: The information is screened once for the rapid miner tool, and the set of frequent items, F, is then calculated and arranged in the size of the items with support values using FP-growth, as shown in Table III.



TABLE III. GENERATE A FREQUENT ITEMSET USING FP-GROWTH

Size	Support	Item 1	Item 2	Item 3
1	1	browser:access:UNKNOWN	-	-
1	1	enrollment_id	-	-
1	1	server:problem:problem	-	-
1	0.727	server:access:UNKNOWN	-	-
1	0.704	server:navigate:UNKNOWN	-	-
1	0.701	server:discussion:UNKNOWN	-	-
1	0.576	server:access:chapter	-	-
1	0.338	browser:video:video	-	-
1	0.326	browser:access:sequential	-	-
2	1	browser:access:UNKNOWN	enrollment_id	-
2	1	browser:access:UNKNOWN	server:problem:problem	-
2	0.727	browser:access:UNKNOWN	server:access:UNKNOWN	-
2	0.704	browser:access:UNKNOWN	server:navigate:UNKNOWN	-
2	0.701	browser:access:UNKNOWN	server:discussion:UNKNOWN	-
2	0.576	browser:access:UNKNOWN	server:access:chapter	-
2	0.429	browser:access:UNKNOWN	browser:problem:combinedopenended	-
2	0.338	browser:access:UNKNOWN	browser:video:video	-
2	0.326	browser:access:UNKNOWN	browser:access:sequential	-
2	1	enrollment_id	server:problem:problem	-
2	0.727	enrollment_id	server:access:UNKNOWN	-
2	0.704	enrollment_id	server:navigate:UNKNOWN	-
2	0.701	enrollment_id	server:discussion:UNKNOWN	-
2	0.576	enrollment_id	server:access:chapter	-
2	0.429	enrollment_id	browser:problem:combinedopenended	-
2	0.338	enrollment_id	browser:video:video	-
2	0.326	enrollment_id	browser:access:sequential	-
2	0.727	server:problem:problem	server:access:UNKNOWN	-
2	0.704	server:problem:problem	server:navigate:UNKNOWN	-
2	0.7	server:problem:problem	server:discussion:UNKNOWN	-
2	0.576	server:problem:problem	server:access:chapter	-
2	0.429	server:problem:problem	browser:problem:combinedopenended	-
2	0.338	server:problem:problem	browser:video:video	-
2	0.326	server:problem:problem	browser:access:sequential	-
2	0.671	server:access:UNKNOWN	server:navigate:UNKNOWN	-
2	0.672	server:access:UNKNOWN	server:discussion:UNKNOWN	-
2	0.55	server:access:UNKNOWN	server:access:chapter	-
2	0.429	server:access:UNKNOWN	browser:problem:combinedopenended	-
2	0.32	server:access:UNKNOWN	browser:video:video	-
2	0.315	server:access:UNKNOWN	browser:access:sequential	-
2	0.65	server:navigate:UNKNOWN	server:discussion:UNKNOWN	-
2	0.576	server:navigate:UNKNOWN	server:access:chapter	-
2	0.428	server:navigate:UNKNOWN	browser:problem:combinedopenended	-

2	0.302	server:navigate:UNKNOWN	browser:video:video	-
2	0.326	server:navigate:UNKNOWN	browser:access:sequential	-
2	0.57	server:discussion:UNKNOWN	server:access:chapter	-
2	0.414	server:discussion:UNKNOWN	browser:problem:combinedopenended	-
2	0.314	server:discussion:UNKNOWN	browser:video:video	-
2	0.313	server:discussion:UNKNOWN	browser:access:sequential	-
2	0.374	server:access:chapter	browser:problem:combinedopenended	-
3	1	browser:access:UNKNOWN	enrollment_id	server:problem:problem
3	0.727	browser:access:UNKNOWN	enrollment_id	server:access:UNKNOWN
3	0.704	browser:access:UNKNOWN	enrollment_id	server:navigate:UNKNOWN
3	0.701	browser:access:UNKNOWN	enrollment_id	server:discussion:UNKNOWN
3	0.576	browser:access:UNKNOWN	enrollment_id	server:access:chapter
3	0.429	browser:access:UNKNOWN	enrollment_id	browser:problem:combinedopenended
3	0.338	browser:access:UNKNOWN	enrollment_id	browser:video:video
3	0.326	browser:access:UNKNOWN	enrollment_id	browser:access:sequential
3	0.727	browser:access:UNKNOWN	server:problem:problem	server:access:UNKNOWN
3	0.704	browser:access:UNKNOWN	server:problem:problem	server:navigate:UNKNOWN
3	0.7	browser:access:UNKNOWN	server:problem:problem	server:discussion:UNKNOWN
3	0.576	browser:access:UNKNOWN	server:problem:problem	server:access:chapter
3	0.429	browser:access:UNKNOWN	server:problem:problem	browser:problem:combinedopenended
3	0.338	browser:access:UNKNOWN	server:problem:problem	browser:video:video
3	0.326	browser:access:UNKNOWN	server:problem:problem	browser:access:sequential
3	0.671	browser:access:UNKNOWN	server:access:UNKNOWN	server:navigate:UNKNOWN
3	0.672	browser:access:UNKNOWN	server:access:UNKNOWN	server:discussion:UNKNOWN
3	0.55	browser:access:UNKNOWN	server:access:UNKNOWN	server:access:chapter
3	0.429	browser:access:UNKNOWN	server:access:UNKNOWN	browser:problem:combinedopenended
3	0.32	browser:access:UNKNOWN	server:access:UNKNOWN	browser:video:video
3	0.315	browser:access:UNKNOWN	server:access:UNKNOWN	browser:access:sequential
3	0.65	browser:access:UNKNOWN	server:navigate:UNKNOWN	server:discussion:UNKNOWN
3	0.576	browser:access:UNKNOWN	server:navigate:UNKNOWN	server:access:chapter
3	0.428	browser:access:UNKNOWN	server:navigate:UNKNOWN	browser:problem:combinedopenended
3	0.57	browser:access:UNKNOWN	server:discussion:UNKNOWN	server:access:chapter
3	0.414	browser:access:UNKNOWN	server:discussion:UNKNOWN	browser:problem:combinedopenended
3	0.314	browser:access:UNKNOWN	server:discussion:UNKNOWN	browser:video:video
3	0.313	browser:access:UNKNOWN	server:discussion:UNKNOWN	browser:access:sequential
3	0.374	browser:access:UNKNOWN	server:access:chapter	browser:problem:combinedopenended
3	0.727	enrollment_id	server:problem:problem	server:access:UNKNOWN
3	0.704	enrollment_id	server:problem:problem	server:navigate:UNKNOWN
3	0.7	enrollment_id	server:problem:problem	server:discussion:UNKNOWN
3	0.576	enrollment_id	server:problem:problem	server:access:chapter
3	0.429	enrollment_id	server:problem:problem	browser:problem:combinedopenended
3	0.338	enrollment_id	server:problem:problem	browser:video:video
3	0.326	enrollment_id	server:problem:problem	browser:access:sequential
3	0.671	enrollment_id	server:access:UNKNOWN	server:navigate:UNKNOWN

3	0.672	enrollment_id	server:access:UNKNOWN	server:discussion:UNKNOWN
3	0.55	enrollment_id	server:access:UNKNOWN	server:access:chapter
3	0.429	enrollment_id	server:access:UNKNOWN	browser:problem:combinedopenended
3	0.32	enrollment_id	server:access:UNKNOWN	browser:video:video
3	0.315	enrollment_id	server:access:UNKNOWN	browser:access:sequential
3	0.65	enrollment_id	server:navigate:UNKNOWN	server:discussion:UNKNOWN
3	0.576	enrollment_id	server:navigate:UNKNOWN	server:access:chapter
3	0.428	enrollment_id	server:navigate:UNKNOWN	browser:problem:combinedopenended
3	0.302	enrollment_id	server:navigate:UNKNOWN	browser:video:video
3	0.326	enrollment_id	server:navigate:UNKNOWN	browser:access:sequential
3	0.57	enrollment_id	server:discussion:UNKNOWN	server:access:chapter
3	0.414	enrollment_id	server:discussion:UNKNOWN	browser:problem:combinedopenended
3	0.314	enrollment_id	server:discussion:UNKNOWN	browser:video:video
3	0.313	enrollment_id	server:discussion:UNKNOWN	browser:access:sequential
3	0.374	enrollment_id	server:access:chapter	browser:problem:combinedopenended
3	0.671	server:problem:problem	server:access:UNKNOWN	server:navigate:UNKNOWN
3	0.672	server:problem:problem	server:access:UNKNOWN	server:discussion:UNKNOWN
3	0.55	server:problem:problem	server:access:UNKNOWN	server:access:chapter
3	0.429	server:problem:problem	server:access:UNKNOWN	browser:problem:combinedopenended
3	0.32	server:problem:problem	server:access:UNKNOWN	browser:video:video
3	0.315	server:problem:problem	server:access:UNKNOWN	browser:access:sequential
3	0.65	server:problem:problem	server:navigate:UNKNOWN	server:discussion:UNKNOWN
3	0.576	server:problem:problem	server:navigate:UNKNOWN	server:access:chapter
3	0.428	server:problem:problem	server:navigate:UNKNOWN	browser:problem:combinedopenended
3	0.302	server:problem:problem	server:navigate:UNKNOWN	browser:video:video
3	0.326	server:problem:problem	server:navigate:UNKNOWN	browser:access:sequential
3	0.569	server:problem:problem	server:discussion:UNKNOWN	server:access:chapter
3	0.414	server:problem:problem	server:discussion:UNKNOWN	browser:problem:combinedopenended
3	0.314	server:problem:problem	server:discussion:UNKNOWN	browser:video:video
3	0.313	server:problem:problem	server:discussion:UNKNOWN	browser:access:sequential
3	0.374	server:problem:problem	server:access:chapter	browser:problem:combinedopenended
3	0.622	server:access:UNKNOWN	server:navigate:UNKNOWN	server:discussion:UNKNOWN
3	0.55	server:access:UNKNOWN	server:navigate:UNKNOWN	server:access:chapter
3	0.428	server:access:UNKNOWN	server:navigate:UNKNOWN	browser:problem:combinedopenended
3	0.315	server:access:UNKNOWN	server:navigate:UNKNOWN	browser:access:sequential
3	0.545	server:access:UNKNOWN	server:discussion:UNKNOWN	server:access:chapter
3	0.414	server:access:UNKNOWN	server:discussion:UNKNOWN	browser:problem:combinedopenended
3	0.306	server:access:UNKNOWN	server:discussion:UNKNOWN	browser:video:video
3	0.302	server:access:UNKNOWN	server:discussion:UNKNOWN	browser:access:sequential
3	0.374	server:access:UNKNOWN	server:access:chapter	browser:problem:combinedopenended
3	0.57	server:navigate:UNKNOWN	server:discussion:UNKNOWN	server:access:chapter
3	0.412	server:navigate:UNKNOWN	server:discussion:UNKNOWN	browser:problem:combinedopenended
3	0.313	server:navigate:UNKNOWN	server:discussion:UNKNOWN	browser:access:sequential
3	0.374	server:navigate:UNKNOWN	server:access:chapter	browser:problem:combinedopenended

After preprocessing, the dataset contains 72,142 tuples, and the model must be built by selecting features with a minimum support value of greater than 0.5 and a maximum itemset size of 3, as shown in Table IV.

In Table V demonstrates the ReLu activation functions, which are present in the hidden and output layers, including the Adam optimization method and a learning rate of 0.2 with a 10 epoch rate, which is applied to ANN approaches.

**B. Evaluation Metrics**

The performance measures are evaluated metrics[33]like accuracy, precision, recall, and F1 – score, Training & Test Score. In this analysis, the class label is used as a binary classification method.

**C. Comparison and Analysis with the baseline Model**

ML approaches are used in the majority of dropout forecasts. The values in Table VI demonstrate the prediction accuracy of the overall learner behavioral features [26], and in Table VII, the selected features by our technique are significantly better than those of the benchmark method, indicating that our feature extraction method is effective.

TABLE IV. SELECTED LEARNER ACTIVITY PARAMETERS

Features	Description
enrollment_id	The learners' unique enrollment number.
server:problem:problem	The number of issues that the server is encountering determines the behavior.
server:access:UNKNOWN	The behavior is defined by the number of requests received from the server.
server:navigate:UNKNOWN	The size of server-based navigations to other regions of the course is used to calculate the behavior.
server:discussion:UNKNOWN	The behavior is determined by the number of users who access the course forum from the server.
browser:access:UNKNOWN	The learner's behavior is defined by the amount of browser accesses obtained.
Output	Label of the dataset

TABLE V. APPLIED CONSTRAINTS

Constraints	Implications
Learning rate	0.2
Epochs	10
Activation function	ReLu
Optimizer	Adam

TABLE VI. USING BASELINE MODELS, CONTRAST THE OVERALL LEARNER BEHAVIORAL RESULTS

S.No	Metrics/ Learning Method	Accuracy	Precision	Recall	F1-score
1	LR	0.78	0.78	1	0.88
2	DT	0.78	0.78	1	0.88
3	RF	0.78	0.78	1	0.88
4	KNN	0.79	0.79	0.99	0.88
5	NB	0.78	0.78	1	0.9

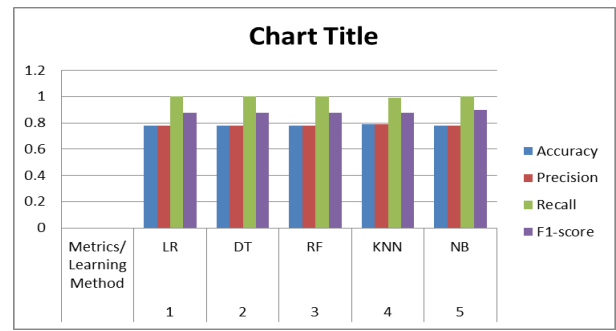


Fig. 3. The Results of Employing Several Classifiers to Predict Overall Learner Behavior.

Fig. 3 shows that the results of several basic supervised machine learning algorithms are used as input data for the whole learner's behavioral actions. In comparison to other models, the KNN model achieves the best accuracy rate of 79%.

To begin, we compared the models to a preset baseline model using the FIAR-ANN model. Table VII shows the findings for the various machine learning algorithms used in the research. The experimental findings with the best values are indicated as strong.

Compare the accuracy level with the selected features by using FIAR-ANN method, entire learner behavior and whole attributes in the dataset as 92%, 78% and 72%. Therefore the vital features produces the best result related to the others.

In terms of the four metric values, the FIAR-ANN model outperforms the other five models. Fig. 4 depicts the evaluation metrics for each model.

TABLE VII. RESULTS IN HYPER PARAMETERS COMPARED TO THE BASELINE MODEL

S.No	Metrics/ Learning Method	Accuracy	Precision	Recall	F1-score
1	LR	0.85	0.86	0.96	0.95
2	DT	0.83	0.84	0.97	0.93
3	RF	0.84	0.86	0.95	0.94
4	KNN	0.86	0.88	0.95	0.95
5	NB	0.84	0.84	0.97	0.8
6	<b>FIAR-ANN</b>	<b>0.92</b>	<b>0.93</b>	<b>0.99</b>	<b>0.91</b>

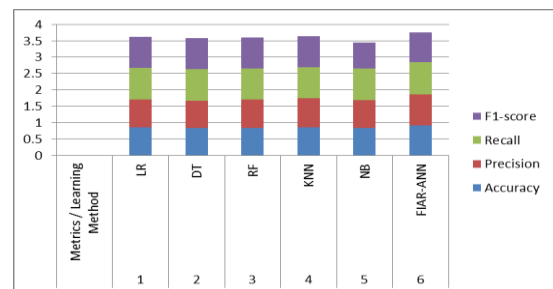


Fig. 4. The Extracted Feature Results are Compared to the Baseline Model's Results.

The findings show that certain machine learning models and the selecting input described in this work are better suited to predicting the dropout problem in MOOCs than others.

According to the findings.

1) In the large-data MOOC, the FIAR-ANN model developed in this research focuses on solving the dropout prediction problem and improving the baseline technique.

2) Using learner behavioral data from the KDD CUP 2015 dataset, the FIAR-ANN model, which has 92 percent accuracy, can be used to predict dropout rates for new programs.

## V. CONCLUSION

Researchers devised a number of methods to predict learner dropout in online programs. From primary behavioral data, we identify and retrieve a number of interpretive behavior aspects. The frequent candidate itemset is generated using an association rule mining-FP growth method. The itemset contains the most often observed learner behavior. Then select the parameters that are present in the three most common items. An artificial neural network approach is applied for the evaluation of the selected parameters. Our proposed method for assessing the efficacy of the KDD CUP 2015 dataset parameters and their values was applied to several machine learning approaches. Then compare the ANN and ML techniques' performance measures. The FIAR-ANN model incorporates the consequences of behavioral characteristics of dropouts, promptly enhancing the dropout prediction accuracy. In this work, we are limited to evaluate learner behavioral activities on the computer related courses, but enhance the work in future with other essential characteristics and also in other discipline courses. Put more emphasis on the attributes that are relevant to different types of platforms in the future.

## REFERENCES

- [1] S. Nithya and S. Umarani, "A Novel Analysis of MOOC's Prediction," vol. 17, no. 10.
- [2] E. W. Winarni, E. P. Purwandari, and S. Hafiza, "Automatic Essay Assessment for Blended Learning in Elementary School," *Int. J. Adv. Sci. Eng. Inf. Technol.*, vol. 12, no. 1, pp. 85–91, 2022, doi: 10.18517/ijaseit.12.1.11835.
- [3] K. Seery, A. Barreda, S. Hein, and J. Hiller, "Retention strategies for online students: A systematic literature review," *J. Glob. Educ. Res.*, vol. 5, no. 1, pp. 72–84, Jun. 2021, doi: 10.5038/2577-509X.5.1.1105.
- [4] Z. Xie, "Bridging MOOC Education and Information Sciences: Empirical Studies," *IEEE Access*, vol. 7, pp. 74206–74216, 2019, doi: 10.1109/ACCESS.2019.2921009.
- [5] S. Ranjeeth, T. P. Latchoumi, and P. V. Paul, "A Survey on Predictive Models of Learning Analytics," *Procedia Comput. Sci.*, vol. 167, no. 2018, pp. 37–46, 2020, doi: 10.1016/j.procs.2020.03.180.
- [6] M. M. Rahman, Y. Watanobe, R. U. Kiran, T. C. Thang, and I. Paik, "Impact of Practical Skills on Academic Performance: A Data-Driven Analysis," *IEEE Access*, vol. 9, pp. 139975–139993, 2021, doi: 10.1109/ACCESS.2021.3119145.
- [7] R. Eusoff, A. M. Zin, and S. M. Salleh, "A Flipped Classroom Framework for Teaching and Learning of Programming," *Int. J. Adv. Sci. Eng. Inf. Technol.*, vol. 12, no. 2, pp. 539–549, 2022, doi: 10.18517/ijaseit.12.2.14909.
- [8] F. Hlioui, N. Aloui, and F. Gargouri, "A withdrawal prediction model of at-risk learners based on behavioural indicators," *Int. J. Web-Based Learn. Teach. Technol.*, vol. 16, no. 2, pp. 32–53, 2021, doi: 10.4018/IJWLTT.2021030103.
- [9] J. Shobana and M. Murali, "Adaptive particle swarm optimization algorithm based long short-term memory networks for sentiment analysis," *J. Intell. Fuzzy Syst.*, vol. 40, no. 6, pp. 10703–10719, 2021, doi: 10.3233/JIFS-201644.
- [10] G. Gavisiddappa, S. Mahadevappa, and C. M. Patil, "Multimodal biometric authentication system using modified relief feature selection and multi support vector machine," *Int. J. Intell. Eng. Syst.*, vol. 13, no. 1, pp. 1–12, 2020, doi: 10.22266/ijies2020.0229.01.
- [11] Y. Zheng, Z. Gao, Y. Wang, and Q. Fu, "MOOC Dropout Prediction Using FWTS-CNN Model Based on Fused Feature Weighting and Time Series," *IEEE Access*, vol. 8, pp. 225324–225335, 2020, doi: 10.1109/ACCESS.2020.3045157.
- [12] K. Sharma, K. Mangaraska, N. van Berkel, M. Giannakos, and V. Kostakos, "Information flow and cognition affect each other: Evidence from digital learning," *Int. J. Hum. Comput. Stud.*, vol. 146, no. September 2020, p. 102549, 2021, doi: 10.1016/j.ijhcs.2020.102549.
- [13] J. Chen, J. Feng, X. Sun, N. Wu, Z. Yang, and S. Chen, "MOOC Dropout Prediction Using a Hybrid Algorithm Based on Decision Tree and Extreme Learning Machine," *Math. Probl. Eng.*, vol. 2019, 2019, doi: 10.1155/2019/8404653.
- [14] C. Jin, "MOOC student dropout prediction model based on learning behavior features and parameter optimization," *Interact. Learn. Environ.*, vol. 0, no. 0, pp. 1–19, 2020, doi: 10.1080/10494820.2020.1802300.
- [15] J. Shobana and M. Murali, "An efficient sentiment analysis methodology based on long short-term memory networks," *Complex Intell. Syst.*, vol. 7, no. 5, pp. 2485–2501, 2021, doi: 10.1007/s40747-021-00436-4.
- [16] A. U. Haq, D. Zhang, H. Peng, and S. U. Rahman, "Combining Multiple Feature-Ranking Techniques and Clustering of Variables for Feature Selection," *IEEE Access*, vol. 7, pp. 151482–151492, 2019, doi: 10.1109/ACCESS.2019.2947701.
- [17] B. Wei, W. Zhang, X. Xia, Y. Zhang, F. Yu, and Z. Zhu, "Efficient Feature Selection Algorithm Based on Particle Swarm Optimization with Learning Memory," *IEEE Access*, vol. 7, pp. 166066–166078, 2019, doi: 10.1109/ACCESS.2019.2953298.
- [18] A. Khan and S. K. Ghosh, "Student performance analysis and prediction in classroom learning: A review of educational data mining studies," vol. 26, no. 1. *Education and Information Technologies*, 2021.
- [19] S. Singh and S. P. Lal, "Using feature selection and association rule mining to evaluate digital courseware," *Int. Conf. ICT Knowl. Eng.*, 2013, doi: 10.1109/ICTKE.2013.6756286.
- [20] A. A. Raweh, M. Nassef, and A. Badr, "A Hybridized Feature Selection and Extraction Approach for Enhancing Cancer Prediction Based on DNA Methylation," *IEEE Access*, vol. 6, pp. 15212–15223, 2018, doi: 10.1109/ACCESS.2018.2812734.
- [21] S. Umarani, T. R. Chaithanya, and M. Divya, "Deployment of P-Cycle in Optical Networks : A Data Mining Approach Deployment of P-Cycle in Optical Networks : A Data Mining Approach," no. September 2016, 2019.
- [22] "KDD CUP 2015 Dataset." <https://data-mining.philippe-fourmieviger.com/the-kddcup-2015-dataset-download-link/>.
- [23] A. Moubayed, M. Injadat, A. Shami, and H. Lutfiyya, "Relationship between student engagement and performance in e-learning environment using association rules," *EDUNINE 2018 - 2nd IEEE World Eng. Educ. Conf. Role Prof. Assoc. Contemp. Eng. Careers Proc.*, pp. 1–6, 2018, doi: 10.1109/EDUNINE.2018.8451005.
- [24] A. Zuiderwijk, Y. Chen, and F. Salem, "Implications of the use of artificial intelligence in public governance: A systematic literature review and a research agenda," *Gov. Inf. Q.*, no. March, p. 101577, 2021, doi: 10.1016/j.giq.2021.101577.
- [25] C. H. Yu, J. Wu, and A. C. Liu, "Predicting learning outcomes with MOOC clickstreams," *Educ. Sci.*, vol. 9, no. 2, 2019, doi: 10.3390/educsci9020104.
- [26] S. Nithya and Dr.S.Umarani, "Comparative Analysis of the Learning on KDD Cup 2015 Dataset," *Webology*, vol. 19, no. 1, pp. 705–717, 2022, doi: 10.14704/web/v19i1/web19050.

- [27] D. Peng and G. Aggarwal, "Modeling MOOC Dropouts," *Entropy*, vol. 10, no. 114, p. 49944, 2015.
- [28] M. Şahin, "A Comparative Analysis of Dropout Prediction in Massive Open Online Courses," *Arab. J. Sci. Eng.*, 2020, doi: 10.1007/s13369-020-05127-9.
- [29] S. Dass, K. Gary, and J. Cunningham, "Predicting student dropout in self-paced mooc course using random forest model," *Inf.*, vol. 12, no. 11, 2021, doi: 10.3390/info12110476.
- [30] J. L. Rastrollo-Guerrero, J. A. Gómez-Pulido, and A. Durán-Domínguez, "Analyzing and predicting students' performance by means of machine learning: A review," *Appl. Sci.*, vol. 10, no. 3, 2020, doi: 10.3390/app10031042.
- [31] D. Romahadi, A. A. Luthfie, W. Suprihatiningsih, and H. Xiong, "Designing Expert System for Centrifugal using Vibration Signal and Bayesian Networks," *Int. J. Adv. Sci. Eng. Inf. Technol.*, vol. 12, no. 1, pp. 23–31, 2022, doi: 10.18517/ijaseit.12.1.12448.
- [32] P. M. Moreno-Marcos, C. Alario-Hoyos, P. J. Muñoz-Merino, and C. D. Kloos, "Prediction in MOOCs: A Review and Future Research Directions," *IEEE Trans. Learn. Technol.*, vol. 12, no. 3, pp. 384–401, 2019, doi: 10.1109/TLT.2018.2856808.
- [33] Y. Wen, Y. Tian, B. Wen, Q. Zhou, G. Cai, and S. Liu, "Consideration of the local correlation of learning behaviors to predict dropouts from MOOCs," *Tsinghua Sci. Technol.*, vol. 25, no. 3, pp. 336–347, 2020, doi: 10.26599/TST.2019.9010013.



# Sentiment Analysis of Online Movie Reviews using Machine Learning

Isaiah Steinke, Justin Wier, Lindsay Simon, Raed Seetan  
Department of Mathematics and Statistics  
Department of Computer Science  
Slippery Rock University  
Slippery Rock, PA, USA

**Abstract**—Many websites encourage their users to write reviews for a wide variety of products and services. In particular, movie reviews may influence the decisions of potential viewers. However, users face the arduous tasks of summarizing the information in multiple reviews and determining the useful and relevant reviews among a very large number of reviews. Therefore, we developed machine learning (ML) models to classify whether an online movie review has positive or negative sentiment. We utilized the Stanford Large Movie Review Dataset to build models using decision trees, random forests, and support vector machines (SVMs). Further, we compiled a new dataset comprising reviews from IMDb posted in 2019 and 2020 to assess whether sentiment changed owing to the coronavirus disease 2019 (COVID-19) pandemic. Our results show that the random forests and SVM models provide the best classification accuracies of 85.27% and 86.18%, respectively. Further, we find that movie reviews became more negative in 2020. However, statistical tests show that this change in sentiment cannot be discerned from our model predictions.

**Keywords**—Decision tree; machine learning (ML); natural language processing (NLP); random forests; sentiment analysis; support vector machine (SVM); reviews

## I. INTRODUCTION

As we move more of our lives online, it becomes possible to find people's thoughts and opinions on nearly anything, from the quality of their commute to their thoughts on the most picayune political issues. One common way that people express their views online is through user reviews. These extend from assessments of mundane household items on Amazon to more commonly reviewed media such as films, music, and video games. Given the growing prevalence of online reviews, researchers have devoted a significant amount of time to determining their sentiment [1]–[3]. The literature related to natural language processing (NLP) is extensive and outlines techniques used to process text and fit models that can determine—among other things—whether a review is positive or negative.

The importance of online reviews has increased in recent years because potential customers are increasingly using them to make purchasing or viewing decisions [1], [3]. However, the number of reviews for many products, movies, and TV shows is rather large, given the massive amount of data available online. As a consequence, summarizing information from multiple reviews and finding reviews that provide relevant and useful information for decision-making are tedious and

formidable tasks for users. Therefore, potential movie viewers could avoid information overload using an automated method that summarizes and classifies the opinions of online movie reviews.

Since the early 2000s, the field of sentiment analysis (or opinion mining) has grown to address many of the challenges associated with determining people's opinions, emotions, and attitudes [1]–[3]. This field has seen explosive growth in the past two decades mainly because of the increased use of social media, the aggregation of reviews on many websites, the prevalence of blogs, and the increased use and development of machine learning (ML) techniques for NLP. Further, the opinions of consumers have had great value in business, politics, marketing, and public relations [3], as determining the opinions of consumers can potentially result in large monetary savings.

In this study, we utilize sentiment analysis to determine a user's sentiment towards a movie using only that user's review (i.e., document-level analysis). Established in 1990, the Internet Movie Database (IMDb) [4] remains a reliable reference for film aficionados, and popular new releases receive thousands of user reviews. Hence, we use two datasets of IMDb reviews in our analyses. The first dataset is a benchmark dataset and is used to train classification models that determine review sentiment. These models are then applied to a second dataset to analyze user reviews submitted before and during the coronavirus disease 2019 (COVID-19) pandemic. Specifically, we use these results to investigate whether there is a change in review sentiment that may be attributed to the increased stress experienced from lockdowns and the threat of COVID-19 infection.

## II. RELATED WORK

Sentiment analysis is currently a very active area of research, especially given the prevalence of vast amounts of data on the Internet that can be mined for sentiment. Sun *et al.* [2] review a variety of NLP techniques for this purpose. They report that the most popular techniques are Naïve Bayes classifiers, support vector machines (SVMs), latent Dirichlet allocation (LDA), and a variety of neural networks. Further, they discuss the notable preprocessing techniques used for NLP, which include tokenization, word segmentation, part-of-speech (POS) tagging, and parsing. In addition, they review toolkits that are currently available, supervised and

unsupervised techniques, and opinion mining at various levels, e.g., document- and sentence-level opinion mining.

Many of the challenges associated with sentiment analysis have been outlined in the surveys in [1] and [3]. In particular, a review may contain both positive and negative sentiments, but the review itself may be either positive, negative, or neutral. Further, a review may contain many words associated with positive sentiment, but the review may be, for example, negative overall. An early study by Peng *et al.* [5] showed that the use of a human-derived list of keywords to determine sentiment at the document level performed worse than a list of keywords chosen with simple statistics. Moreover, some of the keywords chosen with simple statistics would likely not be chosen by a human to determine sentiment.

Hirschberg and Manning [6] review the latest advances in NLP, focusing on advanced topics such as machine translation, spoken dialogue systems, conversational agents, and machine reading. They also discuss the mining of data on social media, which is a rich area for sentiment analysis. Nadkarni *et al.* [7] present an introduction to NLP as it pertains to the medical field, particularly issues that are related to clinical text. They also provide a brief overview of the techniques preferred for medical NLP, which include SVMs, hidden Markov models (HMMs), and conditional random fields (CRFs), and discuss the importance of N-grams.

Hu and Liu [8] present a set of techniques to mine and summarize consumer reviews of products. They demonstrate that their techniques are able to provide useful feature-based summaries of products sold online. The summarization of movie reviews has also been studied using RapidMiner [9] and in mobile environments [10].

Many researchers have used NLP techniques to analyze the sentiment of movie reviews. Pouransari and Ghili [11] classified two datasets—one with binary class labels and the other with multiple classes—using random forests, SVMs, logistic regression, and recursive neural tensor networks (RNTNs). They propose a new type of RNTN called a low-rank RNTN, which is able to reduce computational costs compared to the standard RNTN while retaining a similar accuracy. Govindarajan [12] utilized naïve Bayes and genetic algorithm (GA) classifiers to classify movie reviews, achieving an accuracy of just over 91% with these two classifiers separately. Moreover, they were able to increase the accuracy to 93.80% using a hybrid naïve Bayes/GA classifier. Khan *et al.* [13] presented a method for the classification and summarization of movie reviews. To improve classification accuracy, their method employs unigrams, bigrams, and trigrams. They were able to achieve high accuracies (~90%) on three different datasets using a naïve Bayes classifier with their proposed method.

Sahu and Ahuja [14] utilize feature extraction and ranking to train classifiers based on decision trees, random forests,  $k$  nearest neighbors (KNN), naïve Bayes, and bagging. They achieved accuracies as high as 88.95% with random forests. Kumar *et al.* [15] use a hybrid feature extraction method to determine the sentiment of IMDb movie reviews. They employed SVM, naïve Bayes, KNN, and maximum entropy

classifiers, realizing an accuracy as high as 83.9% with the maximum entropy classifier.

Many of the techniques used in previous studies require high-performance computational resources, e.g., neural networks. Hence, given our available computational resources, we chose to build models using decision trees, random forests, and SVMs. These techniques were chosen because they are well-established powerful ML techniques with robust performance and prior use by other researchers in the NLP domain. Further, these techniques do not require high-performance computational resources and have achieved some of the best classification accuracies according to literature results.

### III. DATA

#### A. Stanford Large Movie Review Dataset

To build suitable models for the classification of movie reviews, we used the benchmark dataset, “Large Movie Review Dataset,” compiled by researchers at Stanford [16], [17] (we will refer to this dataset as the “Stanford dataset” hereafter). This dataset contains 50,000 reviews from IMDb, which are evenly split into training and test sets, each with 25,000 reviews. Each review is stored in a separate plain text file. In addition, each of the test and training sets are evenly balanced to contain 12,500 positive and negative reviews. Positive reviews are defined as reviews with IMDb ratings of 7 or higher, while negative reviews have ratings of 4 or lower. Neutral reviews were excluded from this dataset. Moreover, the number of reviews was limited to no more than 30 per movie since reviews for the same movie tend to have correlated ratings. Finally, the training and test sets comprise a disjoint set of movies, which avoids the potential for performance gains by “memorizing” movie-specific terms.

#### B. IMDb Dataset

To test for potential changes in sentiment amid the COVID-19 pandemic, we assembled our own dataset of 2,498 IMDb reviews for films released in 2019 and 2020. First, we determined the top 100 films released each year according to popularity by the number of ratings that a film received. We further narrowed this list by removing films released in January through March in 2020, as the effects of COVID-19 were not fully prevalent throughout the U.S. until mid-March. We also removed films that did not receive a U.S. release to mitigate the potential problem of foreign language reviews. After these steps, we selected the top 50 films according to the number of ratings for each year.

Second, we scraped the first page of reviews (25 reviews per page, except for two cases) and the rating assigned by the user; for a review with no rating provided by the user, the rating was input as NA. The final dataset includes 1,249 reviews from 2019 and 1,249 reviews from 2020 with data related to the film’s alphanumeric IMDb code, title, date, average rating, number of ratings, year, and the individual user review and rating.

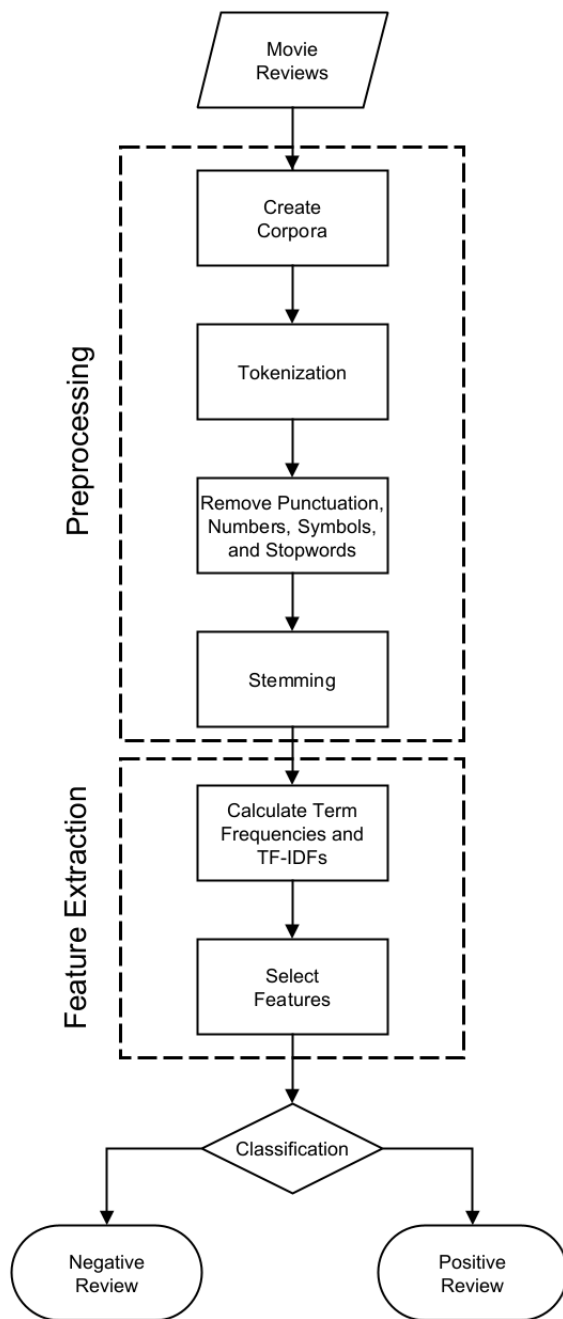


Fig. 1. Flowchart of our Model Process.

#### IV. ANALYSIS METHODS

##### A. Preprocessing

Fig. 1 shows the overall flow of our model process. The raw textual reviews require preprocessing before they can be input into classification models. We utilized the *quanteda* (v. 2.1.2) package in R (v. 4.0.3) to carry out much of this preprocessing. After constructing corpora of reviews, each text review was tokenized, and punctuation, symbols, and numbers were removed. In addition, we removed common English-language stopwords such as “a,” “the,” “of,” and other frequently occurring words that offer little information related

to sentiment. We then converted all words to lower case and reduced them to word stems (e.g., “worst” and “worse” both reduce to the “worst” stem). The use of stemming reduces the total number of features, which is helpful for reducing the computational costs associated with model building.

After calculating the term frequencies for each word stem in each review in the corpora, we calculated the term frequency-inverse document frequency (TF-IDF) for each word stem. The TF-IDF weights the term frequency for a word stem by the inverse of its document frequency in a corpus. Thus, a word stem with a high term frequency that appears in many reviews will have a lower TF-IDF, as it will not be of much use in distinguishing sentiment since it appears in many reviews.

Initially, we built a simple decision tree model with the top 2,500 features having the highest TF-IDFs. The resulting decision tree was primarily built from nine features, all of which were ranked in the top 250 features according to TF-IDF. As more complex models require considerable computation times, we reduced the dataset to use the top 1,000 features with the highest TF-IDFs. Thus, our full training set contained 25,000 reviews with 1,000 features as the input to our models.

The IMDb dataset was preprocessed in the same way as the Stanford dataset so that it used the same 1,000 features/word stems. We also created two versions of the IMDb data. One, which we will refer to as the “unfiltered” version, includes all 2,498 observations/reviews. This means that reviews with ratings of 5 or 6 and those labeled NA were retained. The other, which we will refer to as the “filtered” version, was processed in a similar manner as the Stanford dataset. That is, reviews with ratings of 5, 6, or NA were removed. This reduced the number of reviews to 1,063 for the 2019 films and 1,048 for the 2020 films.

In addition to the 1,000 features, we also explored the use of an additional feature: the length of (or total number of words in) a review. However, after comparing histograms of this feature for the negative and positive reviews in both the training and test sets of the Stanford dataset, we found that the central values and distributions were quite similar. Hence, we did not incorporate this feature since it did not appear that the lengths of reviews would be an effective discriminator of negative and positive reviews.

##### B. Exploratory Analysis of the IMDb Dataset

In sentiment analysis, visualization of the data is often beneficial for understanding the results. After preprocessing all of the textual data in the IMDb dataset, a list of the top 1,000 features according to TF-IDF was reclassified for visualization. By cross-referencing the term frequencies within positive, negative, and neutral reviews, each of the top 1,000 words was assigned a corresponding class. We then used the R package *wordcloud* (v. 2.6) to visualize the 200 most dominant features, as shown in Fig. 2. At a glance, we see that common word stems such as “charact” and “feel” fall into the neutral category and have little use in determining the sentiment of a review. Word stems such as “bad,” “noth,” and “dont” dominate the

negative class, and “good,” “great,” and “watch” stand out for the positive class.

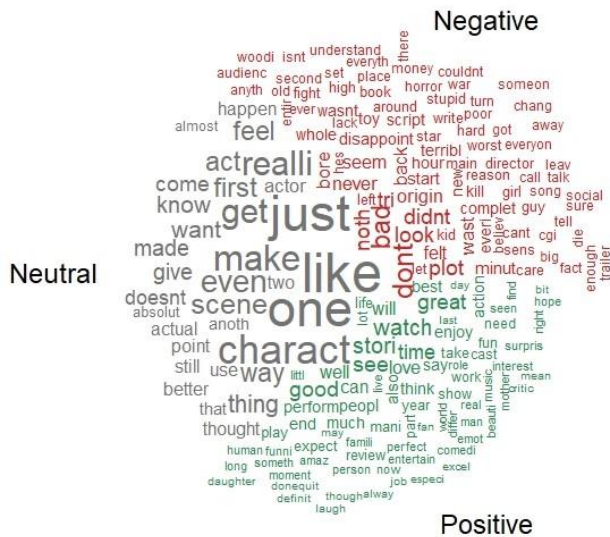


Fig. 2. Word Cloud of the 200 most Prevalent Terms in the IMDb Dataset Categorized as Negative, Neutral, and Positive.

Fig. 3(a) and 3(b) show histograms of the ratings of the reviews in the IMDb dataset for movies released in 2019 and 2020, respectively. Note that movies with no rating (i.e., “NA”) are not accounted for in these histograms. This is a rather small number of ratings: 29 and 28 for 2019 and 2020, respectively. For both years, the numbers of reviews at the extreme ends of the scale, i.e., ratings of 1 and 10, tend to be the highest. Moreover, there is a notable decrease in the number of reviews with a rating of 10 from 2019 to 2020 (with a nearly corresponding increase in movies with a rating of 1).

C. Modeling

All modeling and analyses were carried out in R (v. 4.0.3) on a computer equipped with an AMD Ryzen 7 3800X processor (operating at 3.9/4.5 GHz base/boost clocks) and 32 GB of RAM. We built models using decision trees, random forests, and SVMs, which were respectively implemented using the rpart (v. 4.1-15), ranger (v. 0.12.1), and e1071 (v. 1.7-6) packages in R. In particular, the ranger package allows trees to be built in parallel, which provides a considerable reduction in runtime when building the models.

Since the size of the training set is sufficiently large (25,000 observations), the full training set was split at a ratio of 80:20 into training and validation sets. Given the runtimes for building the random forests and SVM models and our time constraints, we decided against the use of cross-validation (CV) and instead used a single validation set. Models were built using the reduced training set containing 20,000 observations. The validation set containing 5,000 observations was used to tune the hyperparameters of the random forests and SVM models. For the decision tree model, no parameters needed to be tuned.

For the Stanford dataset, we have a binary classification problem, i.e., a review to be classified as either positive or negative. After building a model on the reduced training set,

we tested its performance on the validation set by comparing the class predicted by the model to the actual class. These results are typically summarized with a confusion matrix, from which the numbers of true positives *TP*, true negatives *TN*, false positives *FP*, and false negatives *FN* are obtained. Using these quantities, we assessed the performance of our models using the accuracy *Acc*, precision *Pre*, recall *Rec*, and *F*<sub>1</sub>-score (hereafter denoted by *F*<sub>1</sub>), which are expressed as

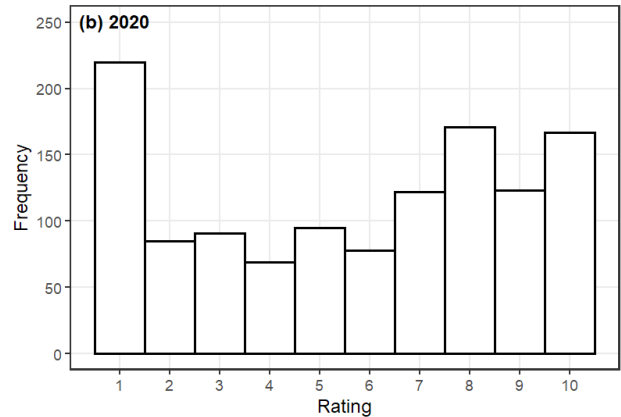
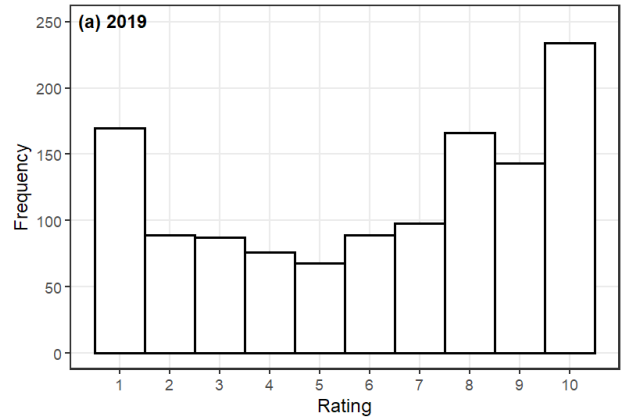


Fig. 3. Histograms of the Ratings of Reviews in the IMDb Dataset for Movies Released in (a) 2019 and (b) 2020.

$$Acc = \frac{TP+TN}{TP+TN+FP+FN} \tag{1}$$

$$Pre = \frac{TP}{TP+FP} \tag{2}$$

$$Rec = \frac{TP}{TP+FN} \tag{3}$$

$$F_1 = \frac{2 \times Pre \times Rec}{Pre + Rec} \tag{4}$$

For the random forests model, we tuned the number of trees, *n*<sub>trees</sub>, and the number of predictors sampled at each split, *n*<sub>preds</sub>, using the training and validation sets with a grid search. The number of trees was varied from 100 to 1000, and the number of predictors was varied from 5 to 200. The hyperparameter values that provided the best overall values of *Acc*, *Pre*, *Rec*, and *F*<sub>1</sub> on the validation set were *n*<sub>trees</sub> = 1,000 and *n*<sub>preds</sub> = 5. These values were then used for the random forests model in subsequent analyses. We additionally tuned the minimum number of observations in a node but found that

this parameter did not substantially increase *Acc*; thus, we used the default value.

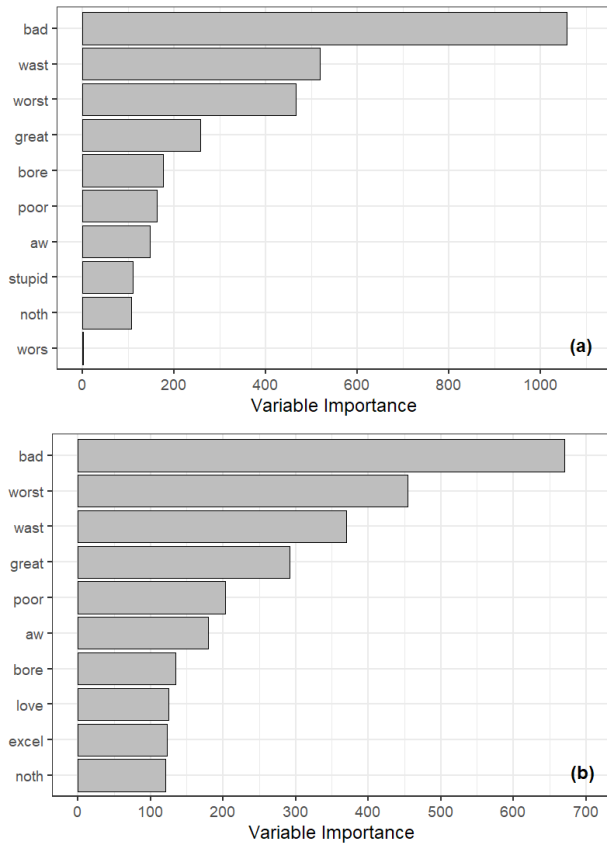


Fig. 4. Plots of the Top 10 most Important Variables for the (a) Decision Tree and (b) Random Forests Models.

For the SVM models, we tuned the hyperparameters for two different kernels: linear and radial. For the linear kernel, we only tuned one parameter, the cost  $C$ , which controls the bias–variance tradeoff [18], over the range of 0.01–1000. For the radial kernel, the hyperparameter  $\gamma$  was also tuned in addition to  $C$ . For this kernel,  $C$  was varied in the range of 1–100, and  $\gamma$  ranged from 0.5 to 5. We found that the radial kernel with  $C = 10$  and  $\gamma = 1$  resulted in the best overall values for *Acc*, *Pre*, *Rec*, and  $F_1$  on the validation set. Again, these values were then used for the SVM model in subsequent analyses. Finally, we also attempted to build SVM models with a polynomial kernel. For the few sets of hyperparameters we tried with this kernel, we obtained poor classification performance on the validation set. For all cases, the models always predicted that a review would be negative, resulting in a low *Acc* value of 0.5. Hence, we did not further tune models with this kernel.

## V. RESULTS

### A. Performance of the Models on the Test Set of the Stanford Dataset

Table I summarizes the performance metrics in (1)–(4) for the decision tree, random forests, and SVM models for the test set of the Stanford dataset. The random forests and SVM models use the best parameters discussed in Section IVC. We

see that the decision tree achieves  $Acc = 73.78\%$ . However, we can greatly increase *Acc* to 85.27% and 86.18% with the random forests and SVM models, respectively. Moreover, the random forests and SVM models provide large increases in *Pre*, *Rec*, and  $F_1$  relative to the decision tree model. Overall, the SVM model has the best overall performance, even though *Pre* is slightly higher for random forests.

TABLE I. PERFORMANCE METRICS FOR THE MODELS ON THE TEST SET OF THE STANFORD DATASET

Model	<i>Acc</i>	<i>Pre</i>	<i>Rec</i>	$F_1$
Decision Tree	73.78%	71.40%	79.34%	75.16%
Random Forests	85.27%	84.93%	85.77%	85.34%
SVM	86.18%	84.62%	88.45%	86.49%

TABLE II. PERCENTAGES OF NEGATIVE AND POSITIVE REVIEWS PREDICTED BY THE MODELS FOR THE UNFILTERED IMDB DATASET

Year	Model	% Negative	% Positive
2019 ( $n = 1249$ )	Decision Tree	35.15%	64.85%
	Random Forests	43.31%	56.69%
	SVM	44.84%	55.16%
2020 ( $n = 1249$ )	Decision Tree	35.23%	64.77%
	Random Forests	46.28%	53.72%
	SVM	48.52%	51.48%

### B. Variable Importance for the Tree-Based Models

For the decision tree model, we noted the variables used in the construction of the tree and their importance. In addition, we looked at the variable importance of the features in the random forests model using the best parameters in Section IVC. Fig. 4(a) and 4(b) show plots of the 10 most important variables for the decision tree and random forests models, respectively.

Both of the decision tree and random forests models found “bad” to be the feature of highest importance. The same eight words appear in the top 10 features of both methods with varying degrees of importance. Further, the majority of these word stems fall into the negative category in the word cloud analysis (e.g., “bad,” “bore,” and “wast”) with some positive word stems (e.g., “great” and “excel”). Notably, these terms make intuitive sense for distinguishing negative and positive movie reviews.

### C. IMDB Dataset

We used our trained models to classify the reviews in the IMDB dataset. Table II summarizes the percentages of movie reviews classified as negative and positive for the unfiltered IMDB dataset. The decision tree model classifies fewer reviews as negative compared to the random forests and SVM models. This is most likely related to the large difference in the performance metrics in Table I, as *Acc* is much worse for the decision tree.

Table III summarizes the percentages of movie reviews classified as negative and positive for the filtered IMDB dataset. Since this dataset has been filtered to only contain negative and positive reviews, as discussed in Section IVA, we have also tabulated the actual percentages of negative and positive reviews using the rating on IMDB for comparison. Here again, we see that the decision tree model classifies fewer reviews as negative compared to the other models. However,

we see that it is the furthest away from the actual percentages than the random forests and SVM models. Surprisingly, the percentages for the random forests model are closer to the actual percentages than those for the SVM model, despite the slightly lower performance metrics for the random forests model. This may be partially attributed to the slightly higher value of *Pre* for the random forests model relative to that of the SVM model. This suggests that it might be better to tune the models to control *FP*.

TABLE III. PERCENTAGES OF NEGATIVE AND POSITIVE REVIEWS PREDICTED BY THE MODELS FOR THE FILTERED IMDB DATASET

Year	Model	% Negative	% Positive
2019 ( <i>n</i> = 1063)	Decision Tree	33.68%	66.32%
	Random Forests	40.73%	59.27%
	SVM	42.80%	57.20%
	Actual	39.70%	60.30%
2020 ( <i>n</i> = 1048)	Decision Tree	33.87%	66.13%
	Random Forests	44.08%	55.92%
	SVM	46.18%	53.82%
	Actual	44.37%	55.63%

The results in Tables II and III show that the percentage of negative reviews increases from 2019 to 2020 for all models and for the actual data. This is also observed in the histograms in Fig. 3. To ascertain whether these results were real, we conducted a chi-squared test of homogeneity [19]. For the unfiltered IMDB data, we found that both of the random forests ( $p = 0.1475$ ) and SVM ( $p = 0.0711$ ) models did not show a statistically significant difference in the distribution of percentages at a level of significance,  $\alpha$ , of 0.05. For the filtered IMDB data, the conclusion was the same for random forests ( $p = 0.1302$ ) and SVM ( $p = 0.1289$ ). However, if we carry out this test with the actual data ( $p = 0.0332$ ), we find that there is a statistically significant difference. Thus, even though there is statistical evidence that there is a change in sentiment from 2019 to 2020 in the actual data, our models do not appear to be sufficiently performant to indicate this difference. This could be an important consideration if we wish to use our models to ascertain changes in sentiment over time, especially if we have unsupervised data, i.e., reviews with no ratings.

We also carried out additional statistical tests of the difference between two population proportions [20] to confirm these results. Here, we only chose to test the SVM model, which has the highest overall performance. Again, there is no statistical evidence (at  $\alpha = 0.05$ ) that there is a difference between population proportions for the unfiltered ( $p = 0.0651$ ) and filtered ( $p = 0.1180$ ) IMDB sets. However, the actual data do indicate that there is statistical evidence of a difference in proportions ( $p = 0.0298$ ), as before.

## VI. DISCUSSION AND CONCLUSION

We have developed models using decision tree, random forests, and SVM ML techniques to classify the sentiment of movie reviews. Our models were trained and built on the Stanford dataset, which is a benchmark dataset in the literature. In addition, we constructed a new dataset consisting of movie reviews from IMDB posted in 2019 and 2020. Using the Stanford dataset, we find that the SVM model provides the best overall performance, with  $Acc = 86.18\%$ ,  $Pre = 84.62\%$ ,

$Rec = 88.45\%$ , and  $F_1 = 86.49\%$ . The random forests model also provides good performance that is slightly worse than the SVM model. The variables that are important for the tree-based models are quite similar, and they tend to utilize words that would be useful in distinguishing negative and positive reviews.

For our IMDB dataset, we find that the random forests and SVM models tend to classify more reviews as negative compared to the decision tree model. Moreover, the random forests model more closely replicates the actual percentages of negative and positive reviews in our filtered IMDB set. However, the models are not able to provide statistical evidence of a change in sentiment from 2019 to 2020. The actual percentages of negative and positive reviews indicate that there is a change in sentiment; that is, issues related to the COVID-19 pandemic may have resulted in an increase in negative reviews. Further time-series analyses may be needed to tell if the increase in negative reviews can be solely attributed to the COVID-19 pandemic. Our analysis does not consider whether this increase was a result of increased negative attitudes elicited by pandemic stressors; studios' decisions to postpone high-profile releases such as *Black Widow*, *Dune*, *F9*, and *No Time to Die*; and other factors that we did not consider.

There is considerable scope for improvement in the performance of our models. During preprocessing, we reduced the number of features to the top 1,000 features according to the TF-IDF. This was done solely to keep the computational times of our models to a reasonable level within our time constraints. It is possible that we may realize additional improvements by increasing the number of features. Further, we did not utilize any feature-reduction techniques, e.g., singular value decomposition (SVD), which may also be employed to increase performance. Although the term frequencies were calculated, we did not build models using them, as longer reviews that repeat a unique term could bias the term frequencies. However, it is possible that models built with the term frequencies could provide better results, especially if we remove terms that appear in both negative and positive reviews at high frequencies. Our time constraints also reduced the number of hyperparameter combinations that we could test and prevented us from employing CV. Additional hyperparameter tuning and CV could increase the performance of our models, although we expect that this is likely to be limited to a few percentage points.

Our models were simple bag-of-words (BoW) models that only used unigrams. As shown by Khan *et al.* [13], we may improve classification performance by including more complex two- and three-word phrases (i.e., bigrams and trigrams, respectively). By using bigrams, a phrase such as "no good," which has a negative connotation, would be used as a bigram instead of the unigrams "no" and "good," which have negative and positive connotations, respectively.

Finally, we were only able to test a small number of ML techniques. According to the literature review in Section II, other techniques can be used to achieve quite high classification accuracies of more than 90%, e.g., naïve Bayes classifiers, GA classifiers, and many types of neural networks.



Hence, additional techniques could be employed in future work to increase classification performance.

#### REFERENCES

- [1] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Found. Trends Inf. Retrieval*, vol. 2, no. 1–2, pp. 1–135, 2008.
- [2] S. Sun, C. Luo, and J. Chen, "A review of natural language processing techniques for opinion mining systems," *Inform. Fusion*, vol. 36, pp. 10–25, 2017.
- [3] B. Liu, "Sentiment analysis and opinion mining," in *Synthesis Lectures on Human Language Technologies*, G. Hirst, Ed. San Rafael, CA, USA: Morgan & Claypool, 2012, pp. 1–167.
- [4] "Internet Movie Database (IMDb)," <http://www.imdb.com> (accessed Apr. 23, 2021).
- [5] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? Sentiment classification using machine learning techniques," in *Proc. Conf. Empirical Methods in Natural Language Processing (EMNLP)*, 2002, pp. 79–86.
- [6] J. Hirschberg and C. D. Manning, "Advances in natural language processing," *Science*, vol. 349, no. 6245, pp. 261–266, Jul. 2015.
- [7] P. M. Nadkarni, L. Ohno-Machado, and W. W. Chapman, "Natural language processing: An introduction," *J. Am. Med. Inform. Assoc.*, vol. 18, pp. 544–551, 2011.
- [8] M. Hu and B. Liu, "Mining and summarizing customer reviews," in *Proc. Tenth ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, Aug. 2004, pp. 168–177.
- [9] A. F. Alsaqer and S. Sasi, "Movie review summarization and sentiment analysis using RapidMiner," in *2017 Int. Conf. Networks & Advances in Computational Technologies (NetACT)*, pp. 329–335.
- [10] C.-L. Liu, W.-H. Hsaio, C.-H. Lee, G.-C. Lu, and E. Jou, "Movie rating and review summarization in mobile environment," *IEEE Trans. Syst., Man, Cybern. Syst. C, Appl. Reviews*, vol. 42, no. 3, pp. 397–407, May 2012.
- [11] H. Pouransari and S. Ghili, "Deep learning for sentiment analysis of movie reviews," 2015. [Online]. Available: <https://cs224d.stanford.edu/reports/PouransariHadi.pdf>.
- [12] M. Govindarajan, "Sentiment analysis of movie reviews using hybrid method of naive Bayes and genetic algorithm," *Int. J. Adv. Comput. Res.*, vol. 3, no. 4, pp.139–145, Dec. 2013.
- [13] A. Khan *et al.* "Summarizing online movie reviews: A machine learning approach to big data analytics," *Sci. Program.*, vol. 2020, Art. no. 5812715, Aug. 2020.
- [14] T. P. Sahu and S. Ahuja, "Sentiment analysis of movie reviews: A study on feature selection & classification algorithms," in *2016 Int. Conf. Microelectronics, Computing and Communications (MicroCom)*.
- [15] H. M. Keerthi Kumar, B. S. Harish, and H. K. Darshan, "Sentiment analysis on IMDb movie reviews using hybrid feature extraction method," *Int. J. Interact. Multimedia Artif. Intell.*, vol. 5, no. 5, pp. 109–114, 2018.
- [16] A. Maas, 2011, "Large Movie Review Dataset," Stanford AI Lab. [Online]. Available: <http://ai.stanford.edu/~amaas/data/sentiment/>.
- [17] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, "Learning word vectors for sentiment analysis," in *49th Annu. Meeting Association for Computational Linguistics (ACL 2011)*.
- [18] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning with Applications in R*, 1st ed. New York, NY, USA: Springer, 2013, ch. 9.
- [19] R. L. Ott & M. Longnecker, *An Introduction to Statistical Methods & Data Analysis*, 7th ed. Boston, MA, USA: Cengage Learning, 2016, sec. 10.5.
- [20] R. L. Ott & M. Longnecker, *An Introduction to Statistical Methods & Data Analysis*, 7th ed. Boston, MA, USA: Cengage Learning, 2016, sec. 10.3.

# Detection and Extraction of Faces and Text Lower Third Techniques for an Audiovisual Archive System using Machine Learning

Khalid El Fayq<sup>1</sup>, Said Tkatek<sup>2</sup>, Lahcen Idouglid<sup>3</sup>, Jaafar Abouchabaka<sup>4</sup>  
LaRIT, Faculty of Sciences, Ibn Tofail University, Kenitra, Morocco

**Abstract**—As part of the audiovisual archive digitization project, which has become a complex field that requires human and material resources, and its automation and optimization have so far represented a center of interest for researchers and media manufacturers, in particular those linked to the integration of artificial intelligence tools in the industry, an elaborate work for the development of an optical character and face recognition model, to digitize the tasks of audiovisual archivist from the manuscript method in automation, from a TV news video. In this article, an approach to develop an example of lower third in Arabic language and facial detection and recognition for news presenter that provide accurate classification results as well as the presentation of different methods and algorithms for Arabic characters. Many studies have been presented in this area, however a satisfactory classification accuracy is yet to be achieved. The comparative state-of-the-art results adopt the latest approaches to study face recognition or OCR, but this model supports both at the same time. It will present the context of realization, the method proposed to extract the texts in the video, using machine learning, about the specificity of the Arabic language, and finally the reasons that govern the decisions taken in the steps of realization. The best results from this approach in real project at the media station was 90.60%. The dataset collected via presenters images and the character dataset via the Pytesseract library.

**Keywords**—Image processing; OpenCV; Tesseract; video OCR; face detection

## I. INTRODUCTION

In recent years, significant and rapid developments in the field of artificial intelligence (AI), the demand for smart applications has increased and found significant interest and use in many fields. As expected, audiovisual production is no exception to this, but radical transformations have taken place in the production process, facial recognition, text, and sound, and also video editing thanks to a set of tools that rely on AI technologies, as a support to the human element, most notably in tasks that require time and repetitive effort. This success in the use of AI is due to two main reasons:

- 1) Big Data Availability, such as photos and videos, in media stations and on the Internet.
- 2) Advances in digital computing manufacturing.

This article is the result of work carried out within Moroccan Television, for the purpose of researching a processing model that will allow the analysis of audiovisual

streams of television news, by analyzing the news presenter faces and the writing in each news coverage.

Videos have become a great source of information; the text of the video contains a substantial amount of information in a non-editable form. If this text is converted into an editable form, it becomes easier and more efficient to store and redistribute [1].

It has been observed that media channels in general have a growing need for automatic facial and handwriting detection systems to integrate into their systems.

One only has to think of the need for the development of active systems to extract the data in the television news videos in order to help the archiving service to fill in the current file, which is currently manually filled in on a daily basis in a register (Fig. 1).

The user must first provide the video as the input from which he wants to extract text. The system will then process the video and generate the editable text output. The latter must ensure the reliability of the information to an acceptable percentage. The solution must extract information from people who appear on the screen, using two methods: facial and text recognition in the video.

It should be noted that there are now commercial automatic tools for processing texts, photos, and videos [2][3], with a complex background, alignment and color variants, etc.

This work will be limited to automating the task of gender recognition from faces only and optical character recognition from the Lower Third present in the video. Also proposed in this work is a video processing chain which includes parts of detection, and tracking of faces and text at the same time.

Effective and efficient text extraction has been a difficult topic in recent years, and the Arabic language is one of the most popular languages in the world. Hundreds of millions of people in many countries around the world speak Arabic as their native language. However, due to the complexity of the Arabic language due to its cursive nature, the recognition of printed and handwritten Arabic text remained untouched for a very long time and did not receive the same attention, compared to the Latin script [4].

The remaining of this paper is structured as follows:

related works developed for text and face recognition from scenes/videos, diagnosis and analysis of the implementation

project and its difficulties, in the face detection and Arabic writing step are discussed in Section II, the presentation of Lower third and face extraction technique and the algorithm to follow for the deployment of this system, discusses the process of preparing the data set are mentioned in Section III and Section IV, the conclusion and some future work discussions are mentioned in Section V.

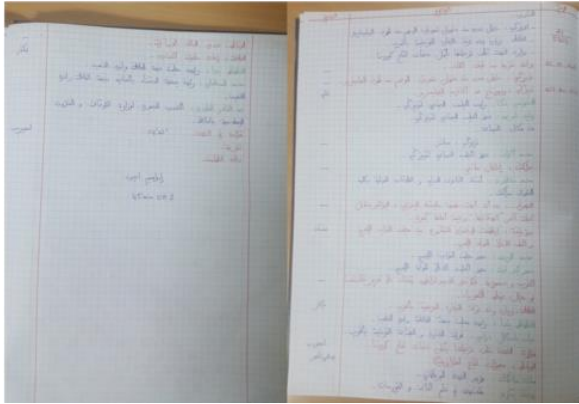


Fig. 1. Handwritten News Archive.

## II. RELATED WORK

Although a large number of printed Arabic character recognition approaches have been proposed in recent years, there is still a need to improve the recognition rate in Arabic OCR systems. This section presents some of these approaches.

Text recognition methods can be categorized into three broad categories, where some approaches recognize text using text segmentation and offering profiling learning with their own features. Methods in the second category recognize text without text segmentation, using a framework based on multiple hypotheses. The methods of the third category improve the text to increase the recognition rate by using binarization of scene images.

Each of these three categories has its own limitations. Approaches in the first category only work well for data from specific scripts, as they need training from their own samples and a classifier to recognize text based on that training. Methods in the second category require multiple hypotheses to set thresholds, but it is unclear how to derive different hypotheses to set specific thresholds. However, the methods of the third category do not need any classifier or hypotheses to define certain thresholds and they also improve the recognition rate through binarization. However, the approaches due to the third category do not provide satisfactory recognition performance for low resolution scene/video images [5], however, these methods perform well on horizontal scene text.

### A. BACKGROUND: Text Recognition

Many systems have been developed to detect text in videos. Each system is based on a specific method and has its associated shortcomings. Some of the commonly used methods to detect text are:

#### 1) Method based on the Sliding Window:

This approach uses a sliding window to search for specific text. It first takes a small rectangular block of a given image.

The rectangular patch has a specific size. Drag this rectangular block over the entire area covered by the image to check if there is any text in this image block. Different sliding window classifiers are used to determine if there is text in the patch. The window is initially placed in the upper left corner of the image, and slides to different positions of the image starting from the first row, then slides through the other rows of the image. This method is slow because the image must be processed at several scales.

#### 2) Method based on Connected Components:

In the connected component-based approach, it first extracts regions of pixels that have a similar color, edge strength, or texture and evaluate each of them to be text or not using techniques of machine learning.

The connected component method works well for caption text with single background images, but it doesn't work well for images with a grouped background.

Text recognition in video images is more difficult than that of natural scene images, difficulties video text analysis are due to complex background images, color variations, font size, camera movement, etc.

Aasim Zafar; Arshad Iqbal [6] compared K-Nearest Neighbor (KNN) and Support Vector Machine (SVM) classifiers in the recognition of printed Arabic characters.

First, feature extraction techniques such as oriented gradient histogram (HOG) and local binary pattern (LBP) have been applied for feature extraction based on the structure of Arabic handwritten texts. SVM has been found to perform better than KNN. Amara et al. [7] presented Arabic OCR using Support Vector Machines (SVM). Although SVM has proven its effectiveness in different fields among other classification tools, SVM has not been effectively applied to recognize Arabic characters.

Saidane and Garcia [8] proposed a convolutional network-based binarization method for the color text area of video images and its performance depends on the amount of training samples used.

Ahmed H and Mahmoud [9] proposed a small-sized printed Arabic text recognition approach based on the estimation of the Hidden Markov Model (HMM). Although applying a hidden Markov model has some advantages (such as no pre-segmentation), the poor image quality of small printed Arabic text makes it difficult to find accurate model boundaries.

Sarfraz and al. [10] proposed an offline Arabic character recognition system. The proposed system has four steps. In the first step, the text preprocessing step removes the isolated pixel and corrects the drift. A pixel is considered isolated if it has no neighboring pixels. Drift is corrected by rotating the image according to the angle with the greatest number of occurrences between all angles of all line segments between any pair of black pixels in the image. In the second step, line and word segmentation is performed using horizontal and vertical projection. Words are segmented into individual characters by comparing the vertical projection profile with a fixed threshold. The feature space is constructed using the moment invariance technique [11].

Zagoris and al. [12] proposed an approach to differentiate handwritten text from machine-printed text. The text image is segmented into blocks. Each block is represented as a vector of words, which contains local features identified using scale-invariant feature transformation. Based on support vector machines, the proposed approach decides whether the text block is handwritten, machine-printed or noise by comparing its word vector with the codebook.

### B. Background: Reconnaissance Facial

Facial recognition started about twenty years ago, and play an important role in various fields, the OpenCV library [13] is free, used in the field of research, because it provides a large number of important tools (more of 500 functions) in the field of computer vision.

The processing and description of audiovisual content presents several opportunities for description through automation and machine learning. Some methods are entirely based on image and sound content, such as facial recognition and audio indexing; other methods rely on text, either inherent in the digital file or extracted from audio or video.

It will discuss the state of the art of facial recognition then face detection methods have been classified by Yang [14], in four approaches:

- Approach based on recognition.
- Approach based on invariant characteristics.
- Approach based on template matching.
- Approach based on appearance

Image classification can be implemented using various supervised techniques such as Naive Bayes [15], K-Nearest Neighbor (KNN) [16], Support vector machines (SVM) [17] [18] [19], Decision trees [20], Random forests [21], Convolutional Neural Network (CNN) [22] [23] and Recurrent Neural Networks (RNN) [24]. These techniques process and classify images into different classes.

In Eidinger, Roe, & Tal [25], they proposed a method based on two tasks based on face representation with local binary patterns (LBP) and linear SVM with dropout. Dropout-SVM is based on the assimilation of a linear SVM to a single layer of a neural network.

Hassner, [26] used the same classification technique as Eidinger, Roe [25], on the front view, projecting 2D points of interest from the front to the 3D face as a reference. They showed that reconstruction of the face in the forward position can improve the performance of facial recognition tasks, in particular gender recognition.

Recently, AZZOPARDI, [27] also proposed a method based on artificial face extraction representation (cosfire-filters), similar to LBP. These representations are also inputs for the linear SVM.

Levi & Tal, [28] implementation of a convolutional neural network (CNN) with tree convolutional layers and three fully connected layers for the task of estimating age and sex; he was specially trained by Audience. To make predictions, they made

several crops of different sizes around the face. The final forecast is the average forecast value of all these crops.

Moreover, Afifi & Abdelhamed, [29] also applied a method based on local facial features, dividing it into several parts (mouth, eyes, nose). They also include insightful strokes around the face while blurring it. Then use these images to train multiple CNNs.

Other papers have also presented a CNN as Ranjan, M. Patel, & Chellappa, [30] proposed two Hyperface models, a CNN network based on the AlexNet architecture and a residual network based on the ResNet 101 architecture. Models are multitasking and can be trained to perform face detection, POI coordinate prediction, pose estimation, and gender recognition simultaneously.

Wolfshaar, F. Karaaba, & A Wiering [31], use a CNN (ImageNet's BVLC) designed to recognize objects belonging to 20,000 categories. In a separate experiment, two datasets (ColorFeret and Adience) of the face were further trained. Then extract the visual features from the penultimate convolutional layer and use them to train the linear SVM.

Ozbulak, Aytar, & Hazim, [32], showed that transfer learning domain-specific models (such as VGG Face) can perform better than recycled CNN with limited data. They achieved this by comparing GilNet (a shallow benchmark CNN trained on the Adience dataset) with two enhanced deep CNNs (one for the VGGFace face and the other more general Alexnet). Then these two enhanced CNNs are used as descriptor extractors, and these descriptors will become the training data of SVM.

### C. Problem Statement

For the moment, Laâyoune TV does not have a computerized archiving system, neither a desktop application nor a web application, and only uses handwritten recording.

On a daily basis, the archivist proceeds to view each video clip, and writes the visual information of each video, which drew my attention to the design of a model that uses two methods of machine learning, in particular the vision by computer, in order to help the television set up an automated system which does the same work of an archivist to extract the data from the video as it is written in the register, with the aim of minimizing the lead time of the archiving task which consumes about one hour of work daily which will free the archivist to use this time for other more important tasks.

Also depending on the global health situation due to the Corona 19 epidemic [33], the Moroccan national radio and television company complies with government laws that encourage remote work.

The main purpose of a text extraction system is to accept video files, detect the text, extract it, and produce an ASCII file including the text in a format that can be used by other applications.

Text detection is performed in each frame of the video. The rectangles representing the location of the text are followed during their period of appearance to associate the corresponding rectangles in the different frames.

This information is necessary to improve the content of the image, which can be achieved by integrating several rectangles containing the same text. This phase must produce sub-images of a quality in accordance with the prerequisites of an OCR process. Therefore, face recognition of news anchors is used to separate each story from the other.

This system makes it possible to extract several information of the news anchor (Fig. 2), the cities (Fig. 3), the people who speak in the microphone (Fig. 4), and the journalists who work in the reports (Fig. 5), in order to provide a complete document.



Fig. 2. Example of a TV News Presenter.



Fig. 3. Example of a City Name in a Report.



Fig. 4. Person Speaks with a Microphone.



Fig. 5. Example of a Team that Produced a Report.

The detection and recognition system must be able to observe a scene. The acquisition conditions of each sequence of images obtained are checked. Usually, capturing the news anchor's face images will be done in front view and in best Full HD (1980x1080) image quality. "Table I" summarizes the video resolution specifications.

TABLE I. VIDEO TECHNICAL SPECIFICATIONS

Type	Video
Codec	MPEG-2 Video (mp2v)
Resolution	1920X1080
Frame rate	25
Planar decoded format	4:2:2 YUV

This system tries to extract and process much information to extract each text frame and recognize faces at the same time (Fig. 6).

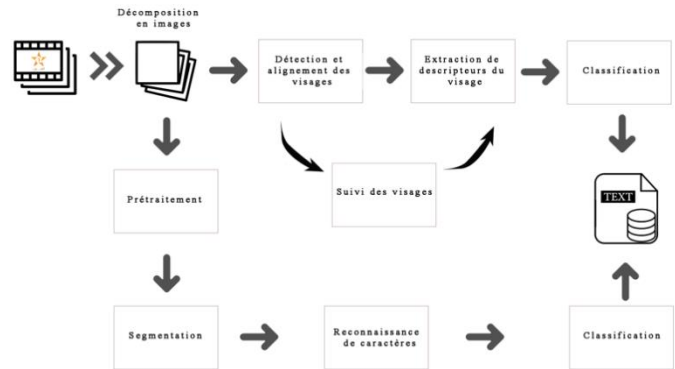


Fig. 6. Processing Model Diagram.

There are two types of text in a video:

- Natural text.
- Overlay text.

1) Natural text is the text that appears in the video when it is recorded. These texts are part of the scene where the video is recorded. Example: nameplate number, text on a man's t-shirt, license plate number (Fig. 7).



Fig. 7. The Name Appears in a Drinking Water Cistern.

2) Overlaid text is the text that was not part of the video when it was recorded, but is overlaid to give additional information about that particular scene. Example: text appearing in the Lower third of TV shows (Fig. 8).



Fig. 8. Example of Lower Third.

Natural text is not of much use because it contains less important information, but overlay text contains information that is of great importance. Therefore, the main objective of the



proposed system is to detect the superimposed text appearing in the video.

News Lower third is static or scrolling depending on the choice of news channel. Both variants have their own challenges. Static Lower third have a fixed space on the video frame to accommodate all text. News channels decrease font size to accommodate more text in some cases. Some channels also perform horizontal compression of text within the frame.

In the context of news videos, the space allocated to the Lower third is fixed and generally designed to accommodate most ligatures, but some are complex.

#### D. Material and Method

Lower third Extraction, has been refined an implementation of the efficient and accurate scene text detector (Tesseract) using the Python programming language [34], on a video to be able to return the bounding boxes of all text in a frame image (Fig. 9).



Fig. 9. Extraction using a Text Detector (Tesseract).

Python has very strong community support with many useful packages and libraries. It is also one of the most popular programming languages for data science, machine learning, artificial intelligence, and scientific computing in general.

The Tesseract engine supports multilingual text recognition [35] [36]. However, recognizing cursive scripts using Tesseract is a difficult task, the Tesseract engine is analyzed and modified for recognition of Arabic writing style. The original Tesseract system has accuracies of 65.59% and 65.84% for 14 and 16 font sizes respectively, while the modified system, with reduced search space, yields accuracies of 97.87% respectively, and 97.71%, this algorithm uses the characteristic of the densities of special symbols in each line of text, which is calculated using the built-in character classifier in Tesseract [37].

### III. TEXT RECOGNITION AND EXTRACTION

#### A. Segmentation

This step is applied to each Frame or image, using the OCR algorithm, the text box is detected. The detected text regions are then refined to increase the efficiency of text extraction. The effectiveness of text detection depends on the font color, text size, background color, and video resolution.

#### B. Classification

Classification is a form of data analysis that extracts patterns describing classes of data, these patterns are called classifiers, and they predict class labels. In this step, the system

must make a decision based on the used algorithm. This analysis can help us better understand the data set.

#### C. Competition Dataset

In this section, he discusses in detail the process of preparing the dataset and the rigorous measures that were taken to ensure maximum diversity throughout in the dataset, through the collection of newsreel video recordings from the source of the National Radio and Television Company (SNRT), in High Definition (HD). At the face recognition level, there is a Database (Dataset) of 7 TV news anchors (Fig. 10), where the system will process these 7 faces to detect them in the video.

The storage of the extracted information for each video is ensured in a separate file bearing the title of the video in the form of a text file, with a small volume.

The videos are stored in a storage server, each video bears the name of the log “newsfeed” + the date (Fig. 11).

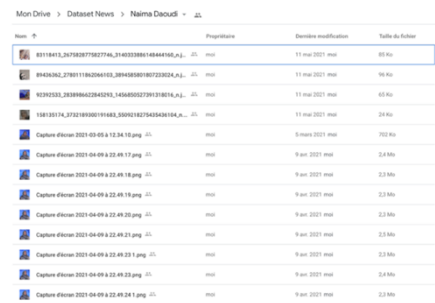


Fig. 10. Training Dataset for Face Detection.

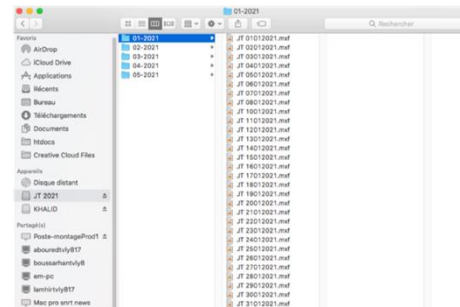


Fig. 11. News Video Clips Stored in Hard Drive.

The advantages of this system:

- Facilitates the task of the archivist and makes the data search and classification simpler.
- Generates a complete newscast information file for extraction. The disadvantage of this system
- Must have a robust computer capable of mass computing and processing information.
- Very slow extraction due to processing each frame for text and face recognition at the same time.

### IV. RESULTS AND DISCUSSION

The extraction of the Lower third was however more complex. It is useful to know the fixed position and speed of the Lower third output for each part. Using this information,



the algorithm extracts the two parts of the recognition to prepare each frame received from the video file in MXF format for the news anchor face localization stage (Fig. 12).

Start of the video with the main presenter (start of the report 0) -----> report 0 -----> senior reporter  
 start report 1 -----> report 1 -----> senior reporter  
 start report 2 -----> report 2 ----->  
 .  
 .  
 .  
 senior reporter start report n -----> report n ----->  
 -----> start the conclusion (end of images) ----->  
 conclusion -----> end

After entering the correct addresses (Fig. 13), you will have access to the reception platform which consists of:

- The "Choose a file" button to load the selected video into a disk.
- The "Download video" button to start the extraction.

In this Application, two techniques have been implemented; face detection and text detection.

Before launching data extraction, it is first necessary in the first time to use this system, to train the news presenters in order to generate a PICKLE type file.

In this case, 900 photos from these seven news presenters are used.

Once the news presenters to be recognized does not correspond to any person in the file (PICKLE), a message will be displayed to indicate that this person is unknown, this message will also appear in the final result file and also a copy of the images in the "image" folder. If the news anchor you want to recognize exists in database, the application will display his first and last name.

The objective of this final report is to filter useful information for the archiving service in order to recognize the news anchor, the people, the number of reports and the employees who produced the television news.

Given the amount of potential software in the audiovisual field that can be based on this type of application, this software must meet the requirements of speed and robustness of results.

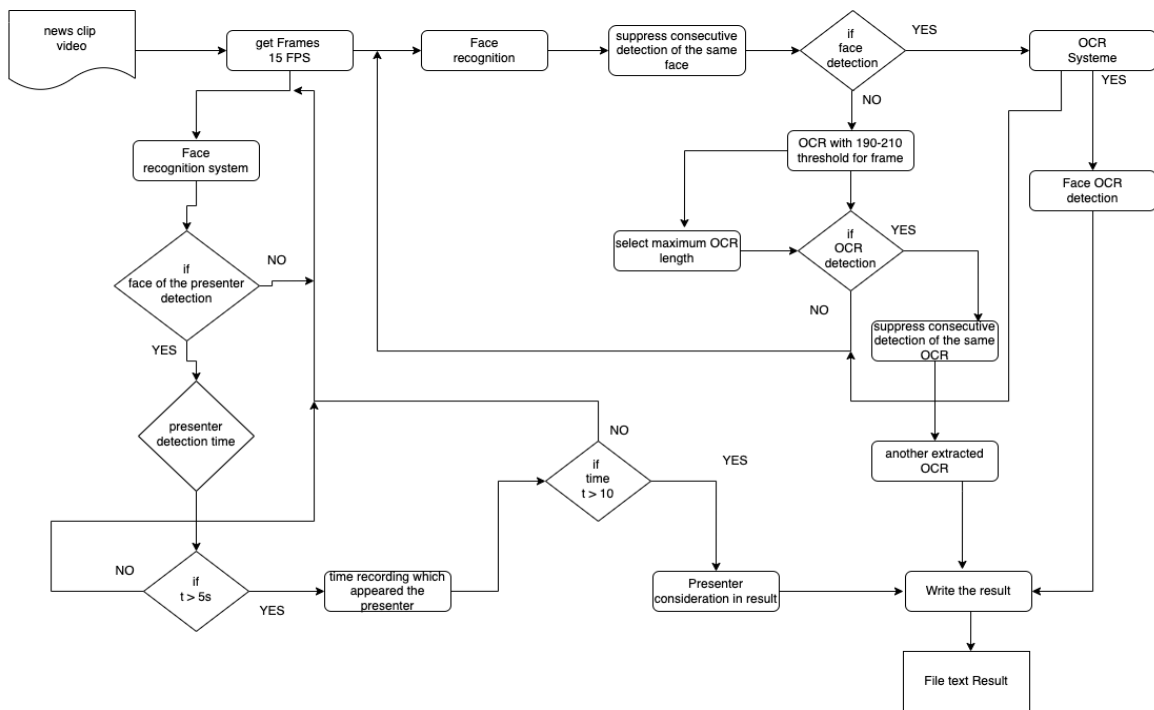


Fig. 12. Facial Recognition and OCR Home Interface.

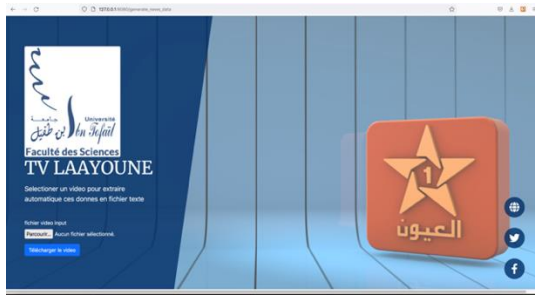


Fig. 13. Facial Recognition and OCR Home Interface.

Finally, a custom web application is created (Fig. 14). The user interface (frontend) is implemented using HTML, CSS and JavaScript. The (backend) of this test application was built using Python with Flask [38], a library for building APIs, combined with the packages and scripts needed to implement OCR and facial recognition. The use of this web application is allowed to get an idea of the video treatments that can be processed successfully.

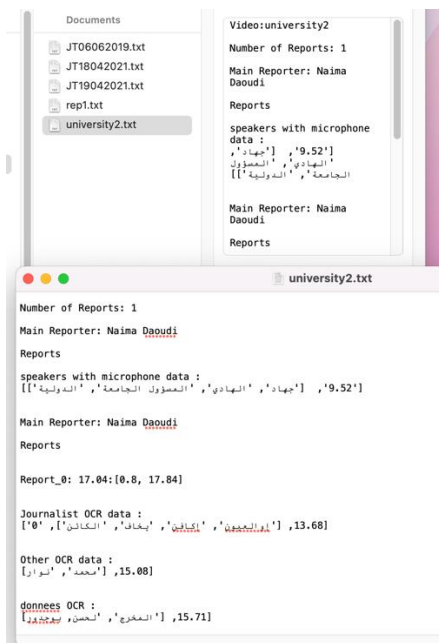


Fig. 14. Extract Result in a System Report.

## V. CONCLUSION

In this paper, a prototype audio-visual archiving system based on face and text detection in Arabic language is designed and implemented. Experimentation shows that the overall recognition accuracy is greater than 88%.

On the other hand, the majority of research works related to video presents a variety of approaches concerning different domains and processes video in general, but does not consider integrating both face and text recognition into a single model. This is therefore the first objective to be attained through this paper.

This approach ensures that human resources help achieve the desired objective by automating the required filing system fixed by managers. This study shows that this problem is very

complex. For this, it was used the algorithm of SVM as well as CNN and library Pytesseract at the same time to solve this problem by the good choice of the parameters and the tolerance coefficient. The optimal solution obtained makes it possible to validate the proposed approach but has a problem of program execution time given the complexity of video processing and computer processors. In future work, he will study similarity semantics for image video text classification. Additionally, it will try to build a bigger model for different languages.

## ACKNOWLEDGMENT

The authors would like to acknowledge the support of the General Administration of the National Society of Moroccan Radio and Television (<http://www.snrt.ma/>), as well as the Faculty of Sciences of Ibn Tofail University (<https://fs.uit.ac.ma/>).

## REFERENCES

- [1] A. A. Shahin, "Printed Arabic Text Recognition using Linear and Nonlinear Regression," 2017. [Online]. Available: [www.ijacsa.thesai.org](http://www.ijacsa.thesai.org)
- [2] A. Mittal, P. P. Roy, P. Singh, and B. Raman, "Rotation and script independent text detection from video frames using sub pixel mapping," *J Vis Commun Image Represent*, vol. 46, pp. 187–198, Jul. 2017, doi: 10.1016/j.jvcir.2017.03.002.
- [3] Josef Chaloupka, A prototype of Audio-Visual Broadcast Transcription System. 2019.
- [4] M. Rashad and N. A. Semary, "CCIS 488 - Isolated Printed Arabic Character Recognition Using KNN and Random Forest Tree Classifiers," 2014.
- [5] A. Kumar Bhunia, G. Kumar, P. Pratim Roy, and R. Balasubramanian, "Text Recognition in Scene Image and Video Frame using Color Channel Selection", doi: 10.1007/s11042-017.
- [6] Institute of Electrical and Electronics Engineers, Institute of Electrical and Electronics Engineers. Delhi Section, and I. INDIACOM (Conference) (14th : 2020 : New Delhi, Machine Reading of Arabic Manuscripts using KNN and SVM Classifiers. 2020.
- [7] M. Amara, K. Zidi, S. Zidi, and K. Ghedira, "CCIS 488 - Arabic Character Recognition Based M-SVM: Review," 2014.
- [8] C. Garcia and Z. Saidane, "Automatic Scene Text Recognition using a Convolutional Neural Network Metric Learning and Siamese Neural Networks View project PhD Thesis-Toward unsupervised activity monitoring with sequence metric learning View project Automatic Scene Text Recognition using a Convolutional Neural Network," 2007. [Online]. Available: <https://www.researchgate.net/publication/251423608>
- [9] Ahmed H. Metwally, Mahmoud I. Khalil, and Hazem M. Abbas, "Offline Arabic handwriting recognition using Hidden Markov Models and Post-Recognition Lexicon Matching," 2017.
- [10] "Optical Character Recognition (OCR) of Arabic Characters," <https://ukdiss.com/examples/optical-character-recognition.php>, 2018.
- [11] J. Chaloupka, "Audio-Visual TV Broadcast Signal Segmentation," in *Advances in Intelligent Systems and Computing*, 2020, vol. 1061, pp. 221–228. doi: 10.1007/978-3-030-31964-9\_21.
- [12] K. Zagoris, I. Pratikakis, A. Antonopoulos, B. Gatos, and N. Papamarkos, "Distinction between handwritten and machine-printed text based on the bag of visual words model," in *Pattern Recognition*, Mar. 2014, vol. 47, no. 3, pp. 1051–1062. doi: 10.1016/j.patcog.2013.09.005.
- [13] Intel corporation, "Open Source Computer Vision Library," 2021.
- [14] M.-H. Yang, D. J. Kriegman, and N. Ahuja, "Detecting Faces in Images: A Survey," 2002.
- [15] S.-C. Hsu, I.-C. Chen, and C.-L. Huang, "Image Classification Using Naive Bayes Classifier With Pairwise Local Observations," XXXX-XXXX, 2017.
- [16] A. Štuliienė and A. Paulauskaitė-Tarasevičienė, "Research on human activity recognition based on image classification methods," 2017.

- [17] C. H. Qian, H. Q. Qiang, and S. R. Gong, "An Image Classification Algorithm Based on SVM," *Applied Mechanics and Materials*, vol. 738–739, pp. 542–545, Mar. 2015, doi: 10.4028/www.scientific.net/amm.738-739.542.
- [18] S. Amassmir, S. Tkatek, O. Abdoun, and J. Abouchabaka, "An intelligent irrigation system based on internet of things to minimize water loss," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 25, no. 1, pp. 504–510, Jan. 2022, doi: 10.11591/ijeecs.v25.i1.pp504-510.
- [19] R. Dahmani, A. Belmzoukia, S. Tkatek, and A. Ait Fora, "Automatic slums identification around normal and smart cities: using Machine-learning on VHR Satellite Imagery," *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 9, no. 5, p. 9071–9079, Oct. 2020, doi: 10.30534/ijatcse/2020/312952020.
- [20] E. Gyimah and D. K. Dake, "Using Decision Tree Classification Algorithm to Predict Learner Typologies for Project-Based Learning," in *Proceedings - 2019 International Conference on Computing, Computational Modelling and Applications, ICCMA 2019*, Mar. 2019, pp. 130–134. doi: 10.1109/ICCMA.2019.00029.
- [21] H. Guan, J. Yu, J. Li, and L. Luo, "RANDOM FORESTS-BASED FEATURE SELECTION FOR LAND-USE CLASSIFICATION USING LIDAR DATA AND ORTHOIMAGERY," 2012.
- [22] A. Sharif, R. H. Azizpour, J. Sullivan, and S. Carlsson, "CNN Features off-the-shelf: an Astounding Baseline for Recognition," 2014.
- [23] S. Dey, R. Bhattacharya, F. Schwenker, and R. Sarkar, "Median filter aided CNN based image Denoising: An ensemble Approach," *Algorithms*, vol. 14, no. 4, Apr. 2021, doi: 10.3390/a14040109.
- [24] J. Chaloupka, K. Palecek, P. Cerva, and J. Zdansky, "Optical character recognition for audio-visual broadcast transcription system," in *11th IEEE International Conference on Cognitive Infocommunications, CogInfoCom 2020 - Proceedings*, Sep. 2020, pp. 229–232. doi: 10.1109/CogInfoCom50765.2020.9237867.
- [25] E. Eidingner, R. Enbar, and T. Hassner, "Age and Gender Estimation of Unfiltered Faces," 2013. [Online]. Available: <http://www.adiance.com>
- [26] T. Hassner, S. Harel, E. Paz, and † Roeenbar, "Effective Face Frontalization in Unconstrained Images," 2015. [Online]. Available: [www.openu.ac.il/home/hassner/projects/frontalize](http://www.openu.ac.il/home/hassner/projects/frontalize)
- [27] G. Azzopardi, A. Greco, A. Saggese, and M. Vento, "Fusion of Domain-Specific and Trainable Features for Gender Recognition from Face Images," *IEEE Access*, vol. 6, pp. 24171–24183, Apr. 2018, doi: 10.1109/ACCESS.2018.2823378.
- [28] G. Levi and T. Hassner, "Age and Gender Classification using Convolutional Neural Networks," 2015. [Online]. Available: [www.openu.ac](http://www.openu.ac).
- [29] M. Afifi and A. Abdelhamed, "AFIF4: Deep Gender Classification based on AdaBoost-based Fusion of Isolated Facial Features and Foggy Faces," Jun. 2017, [Online]. Available: <http://arxiv.org/abs/1706.04277>
- [30] R. Ranjan, V. M. Patel, and R. Chellappa, "HyperFace: A Deep Multi-task Learning Framework for Face Detection, Landmark Localization, Pose Estimation, and Gender Recognition," Mar. 2016, [Online]. Available: <http://arxiv.org/abs/1603.01249>
- [31] Jos van de Wolfshaar, Mahir F. Karaaba, and Marco A. Wiering, *Deep Convolutional Neural Networks and Support Vector Machines for Gender Recognition*. 2015.
- [32] G. Özbülak, Y. Aytar, and H. K. Ekenel, "How transferable are CNN-based features for age and gender classification?," 2016.
- [33] S. Tkatek, A. Belmzoukia, S. Nafai, J. Abouchabaka, and Y. Ibnou-Ratib, "Putting the world back to work: An expert system using big data and artificial intelligence in combating the spread of COVID-19 and similar contagious diseases," *Work*, vol. 67, no. 3, pp. 557–572, 2020, doi: 10.3233/wor-203309.
- [34] <https://www.python.org/>, "Python programming language," [Online], 2021.
- [35] Q. U. A. Akram, S. Hussain, A. Niazi, U. Anjum, and F. Irfan, "Adapting tesseract for complex scripts: An example for Urdu Nastalique," in *Proceedings - 11th IAPR International Workshop on Document Analysis Systems, DAS 2014*, 2014, pp. 191–195. doi: 10.1109/DAS.2014.45.
- [36] R. Smith, "An Overview of the Tesseract OCR Engine," 2007. [Online]. Available: <http://code.google.com/p/tesseract-ocr>.
- [37] Z. Liu and R. Smith, "A simple equation region detector for printed document images in tesseract," in *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR, 2013*, pp. 245–249. doi: 10.1109/ICDAR.2013.56.
- [38] Miguel Grinberg, *Flask Web Development, 2nd Edition*, vol. O'Reilly Media, Inc. 2018.

# Data Recovery Comparative Analysis using Open-based Forensic Tools Source on Linux

Muhammad Fahmi Abdillah<sup>1</sup>

Department of Informatics  
The Islamic University of Indonesia  
Yogyakarta, Indonesia

Yudi Prayudi<sup>2</sup>

Department of Informatics  
The Islamic University of Indonesia  
Yogyakarta, Indonesia

**Abstract**—Data recovery is one of the forensic techniques used to recover data that has been lost or deleted. Data recovery is carried out if there is a condition where the data that has been owned is deleted or damaged. If the data has been lost or deleted or even tampered with, then a forensic expert has several ways to restore data that has been lost or damaged. One of them is to use a complete data recovery method using forensic tools, namely, TSK Recover, FTK Imager, Foremost Recover, and Testdisk Recover. Unfortunately, tools such as FTK imager and TSK recover have a weakness, namely that some damaged or corrupted data files cannot be restored in their entirety; they can only be recovered but not be opened. This study uses a tool comparison method approach using foremost recover and Testdisk recover. It's just that this method cannot be used using the graphic user interface (GUI) but using the CLI (Command Line) in the LINUX operating system. And the files that have been recovered will be fully recovered.

**Keywords**—*Recovery; tools; FTK imager; foremost; Testdisk*

## I. INTRODUCTION

Data loss is a condition where the data that has been owned becomes corrupted or deleted [1]. According to several researchers, there are many companies or individuals who accidentally delete their personal data. It is very important for digital forensic analysts to have the right tools to recover data [2]. All devices store a lot of important data and information that is always used for personal and corporate purposes. Forensic tools are used by thousands of digital forensic professionals. The functionality of forensic tools varies greatly [3].

Currently, there are many simple data recovery tools; several features have been provided consistently for more effective forensic extraction to get the whole data [4], including image storage, file data hashing, data visualization, and data carving on damaged images. However, most of these tools are paid for [5]. Due to the limited inspection features, the extracted data cannot be ported directly to the circuit to extract additional evidence. In this study, I present several tools that will help forensic analysts perform open source-based data recovery on Linux [6].

Data recovery is the process of recovering a problematic or lost system so that it can be recovered as usual [7]. Data recovery is also a forensic technique that is often used to search for digital artifacts that have been lost or deleted from devices such as cellphones, computers, and laptops [8]. Data

backup, which is a preventive measure that is intentionally done to protect data by copying or copying data to other storage media [9].

This study aims to determine the forensic tools that are useful today and in the future. To overcome the occurrence of data loss, a digital forensics expert is needed [10]. Data recovery is one of the techniques that must be mastered by digital forensic experts [11]. If there is data damage or data loss, then it is the job of a forensics officer to recover data that has been lost or damaged [12]. Several cases of data corruption or data loss are one of the challenges that digital forensics experts must face. There are several data recovery tools used by digital forensic experts, such as Autopsy, FTK imager, TSK recover, Foremost, and Testdisk [13].

In the case of previous research, many forensic experts use this tool as a tool to find evidence [14]. This tool is very helpful for recovering data that has been lost or damaged, but this tool has a certain weakness, when restoring data or data recovery, namely data that has been damaged can only be recovered but cannot be opened in its entirety, therefore the solution what is needed is a complete recovery, data that has been retrieved / damaged can be recovered and reopened the same as before. To overcome this problem, a forensic expert uses recovery tools in a storage [15].

Recovery of the data to be recovered is in the allocated space and unallocated space [16]. This space stores all files that are still available and can be read logically, and stores all files that are no longer available, even if they have been deleted from storage and cannot be read logically [17].

From some of the references found, it can be concluded that previous research related to the themes discussed included many case studies that used forensic tools and used several methods to recover lost data [18]. The data is stored in various storage devices such as flash drives, HDDs, SSDs, and RAM. The storage is on mobile devices, computers, and even servers. Data recovery methods also vary depending on the storage to be processed. One of them is using autopsy tools or other forensic tools [19]. This tool is very helpful for forensic experts to find lost data files, such as JPG, MP4, PDF, PNG, Doc, Zip, Rar files, and so on. It's just that this tool has certain weaknesses when it comes to data retrieval or data recovery. Data that has been damaged can only be recovered but cannot be opened in its entirety. Therefore, the solution needed is full

recovery. Data that has been lost or damaged can be recovered and reopened the same as before [20].

Efforts to provide data recovery solutions for handling digital evidence on a storage device such as a smartphone have been discussed by Wilson & Chi (2017) using digital forensic tools to make it easier to acquire data. The most important thing about recovering data is the recovery method because there are many ways to acquire and recover data [12].

However, there are several researchers who provide reviews of data recovery with different techniques and different devices, as discussed by Povar & Bhadran (2011). The carving technique is what is meant. This technique helps in finding hidden or deleted files from digital media or with the data acquisition technique that has been discussed by Jo et al. (2016) regarding data acquisition using forensic tools, namely Autopsy. This technique is very helpful for a forensic expert to collect data or evidence [21].

Several films and photographs in the form of corrupted or damaged JPG, MP4 and PNG files will be used as part of the data to help solve this issue. Additionally, it will be processed afterwards using a number of forensic programs, including Testdisk Recover, Foremost, and the Sleuth Kit Autopsy. These tools will receive this data and begin the recovery procedure [22]. The final step allows for a comparison of numerous efficient tools that can be used to carry out a complete recovery. Using digital forensic tools to aid data capture, Wilson & Chi outline efforts to provide data recovery solutions for digital evidence on a storage device like a smartphone. Since there are numerous ways to obtain and recover data, the recovery method is the most crucial factor [23]. Using a live forensics procedure with the Foremost recovery program or Testdisk recover is the suggested solution for flawless data recovery. Data can be acquired with this tool from storage devices as HDD, SSD, FD, CD/DVD, zip, and rar [12].

This study compares the forensic tool settings that will be utilized to recover deleted or damaged data in the form of data file formats that will be used as evidence in cybercrime case resolution [24]. In this investigation, data recovery is carried out on Linux utilizing the live forensic method. The findings of this study should aid in our understanding of digital forensics, particularly with regard to data recovery [18].

A forensic technique applied while the system is operating is called live forensics. This is because when the system is switched off, the data that needs to be recovered can be lost. This live forensic technique is typically applied to memory scenarios where data can be written to or erased from—this type of memory is also known as volatile memory or non-volatile memory [25].

## II. RESEARCH METHOD

The procedures needed to conduct a study are known as the research technique. These procedures are taken so that a scientific process can be used to tackle the difficulties that develop by providing logical and systematic solutions, as shown in Fig. 1.

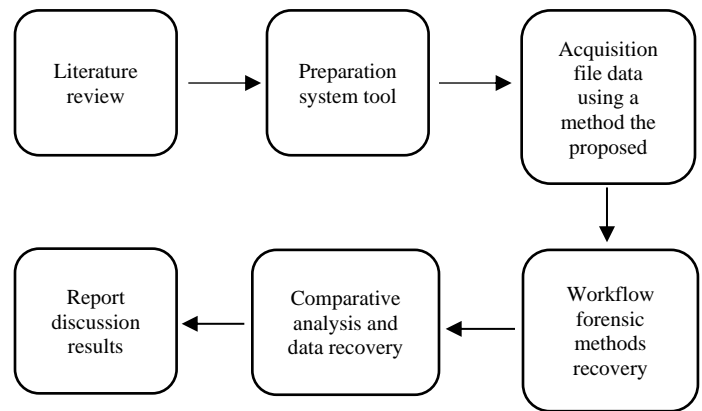


Fig. 1. Research Methodology.

### A. A Study of the Literature

In order to support the overall objective of this research, the purpose of this literature review is to gather information materials on research topics that can be sourced from articles, papers, journals, papers in the form of theories, research reports, or previous findings. We also visit several websites on the internet that is related to these theories about digital forensics, evidence, and recovery.

### B. Tools for System Preparation

This is a step in creating the hardware and software specifications needed for research projects like planning and putting into practice a comparative examination of data recovery utilizing a flash drive. Such as setting up the system and installing software. The physical machine has Microsoft Windows 11 Home installed as its operating system. The employment of physical computer hardware and software as research tools and materials is necessary for the successful operation of the experimental implementation. The following tools and materials are employed in this process:

- MSI Modern 14 laptop with specifications:
  - a. Processor : Intel Core™ I7-10510U CPU  
1.80Ghz
  - b. Memory : 512 GB / 8 GB RAM
  - c. OS : Windows 11 home insider 64-bit
- Flashdisk 8 GB
- TSK recover tool
- Foremost recover tool
- Testdisk Recover tool
- Oracle VM virtual box (CSI Linux)

### C. Proposed Methodology

#### 1) Foremost Recover Forensic Method

In order to replicate the functionality of the DOS carving software for usage on the Linux platform, the most recent recovery technique was developed in March 2001. Special agents Kris Kendall and Jesse Kornblum from the Office of Special Investigations of the US Air Force originally wrote

Formost. The program was altered in 2005 as part of a master's thesis by Nick Mikus, a researcher at the Naval Graduate School's Center for Information Security Studies and Research. Among these changes were improved accuracy and foremost extraction rates.

This method is intended to read and copy data straight from the disk into the computer's memory without taking into account the underlying file system type. The method of file carving is used by Formost Recover to look for header file types that coincide with those in the formost configuration file. There are no alternatives for a graphical user interface, hence the command line interface is primarily used. The JPG, GIF, PNG, BMP, AVI, EXE, MPG, WAV, RIFF, WMV, MOV, PDF, PLE, DOC, ZIP, RAR, HTM, and CPP file formats can all be recovered using the first approach. Additional file types can be specified in the configuration file formost.conf, which is typically located in /usr/local/etc. It can be used to recover data from hard disks that use the ex3, NTFS, or FAT file systems as well as directly from picture files. In example, it can be used to retrieve data from a smartphone via a computer as shown in Fig. 2.

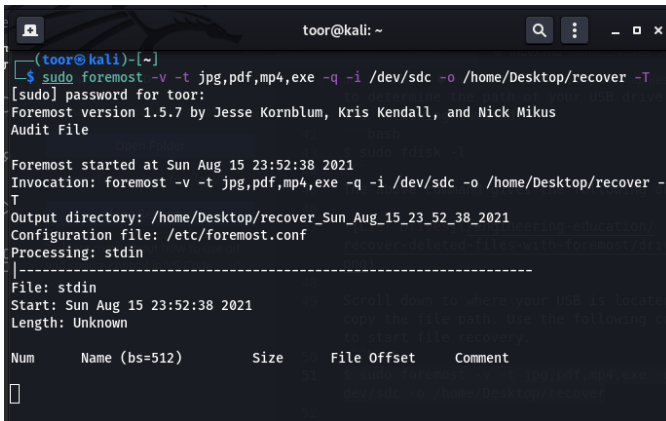


Fig. 2. Example Image Foremost Recover Method.

### 2) Testdisk Recover Forensic Method

A free and open-source utility called Testdisk is used to recover data from missing or deleted partitions. There is no user interface version of this utility; it is CLI driven. A digital forensics specialist can utilize this to restore partitions that are unable to boot due to things like virus attacks and, of course, purposeful or unintentional destruction of the partition table. This testdisk can also do a number of additional tasks, including:

- Recover FAT32 boot sector from backup
- Recover boot sector FAT12/FAT16/FAT32
- Recover NTFS boot sector
- Restore NTFS boot sector from backup
- Fix MFT using MFT mirror
- Find backup superblocks ext2/ext3/ext4
- Undelete file from FAT, exFAT, NTFS, and ext2 file system

- Copy file from FAT, exFAT, NTFS and partitions deleted ext2/ext3/ext4

A forensic expert who is looking into a case involving data loss or corruption will find this Testdisk to be of great assistance. Fig. 3 depicts a sample of the Testdisk recover technique.

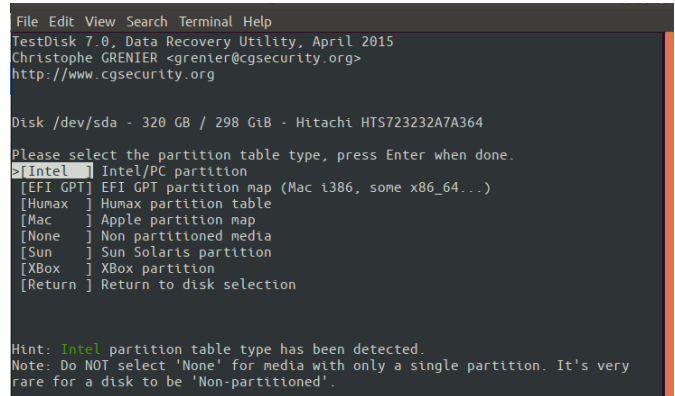


Fig. 3. Example Image Testdisk Recover Method.

### D. Recovery Method Workflow

Workflow recovery methods are phases or steps that digital examiners must go through when performing digital tasks, beginning with preparation, extraction, and analysis. As shown in Fig. 4.

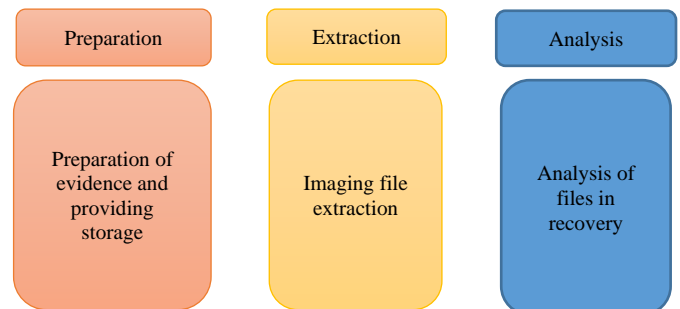


Fig. 4. Workflow Method Recovery.

- Preparation

By providing storage space for data that will be recovered and extracted, we set the stage.

- Extraction

This carries out file extraction by locating and restoring deleted files. The properties of the file structure, deleted data, file name, file size, and location will all be made known through file extraction.

- Analysis

It is in the process of analyzing the outcomes of file-checking. In order to assess or evaluate the success of data file extraction and can suggest the technologies that are best for file recovery in this investigation.



E. Data Recovery Comparative Analysis

Using forensic tools like TSK recover, Formost, and Testdisk recovery, the data recovery stage of comparative analysis was extracted. Damaged data files like JPG, PNG, and MP4 on flash drives, HDDs, SSDs, and other storage will be examined using a data file recovery approach utilizing a number of tools, which will then discover variations that impact data recovery on these tools so that they may be opened again. in full. Tables I and II show many tables that depict the outcomes of the recovery based on the findings of the forensic investigation.

TABLE I. FLASHDISK DATA RECOVERY RESULTS

Storage	Flasdisk	
Tools	TSK recover, FTK imager, Foremost, Testdisk	
Jenis File	JPG, PNG, MP4	
Recovery status	Succeed	Not successful
	√	

TABLE II. HDD/SSD DATA RECOVERY RESULTS

Storage	HHD / SSD	
Tools	TSK recover, FTK imager, Foremost, Testdisk	
Jenis File	JPG, PNG, MP4	
Recovery status	Succeed	Not successful
	√	

III. RESULTS AND DISCUSSIONS

A. Preparation

By providing storage space for the data that will be recovered and extracted, the preparation stage is prepared. In this study, we used flash storage that contained unopenable rar files containing JPG, PNG, and MP4 files. The tests and findings obtained have as their goal getting a full recovery file so you can compare the forensic tools utilized. Some of these files are hashed, as indicated in Tables III and IV, to demonstrate their validity in comparison to the findings of forensic analysis and recovery.

TABLE III. MD5 HASH

No	Data file	Initial MD5 Hash
1	.JPG	459d4d4d38993bb270d9f8d7d5029a5c
2	.PNG	120695b94e5d3bf867862eb42715a4a4
3	.MP4	677f7dca67cdf3741d3f924a668fc2b2

TABLE IV. SHA1 HASH

No	Data file	Initial SHA1 Hash
1	.JPG	1b036089c09444fe5ae1fb0f4279de1f99200fa8
2	.PNG	ccc7286c0ba4a4fb1a61a1793dfa4fc8b60ef60d
3	.MP4	13d1fcd6ef10140dc80726640564aabee08a6161

B. Extraction

File extraction will also expose the characteristics of the file structure, deleted data, file name, file stamps, file size, and location during the extraction stage, which is to extract files by locating and recovering deleted files. The tools that aid in the extraction process run on Linux and use Guymager using the tools TestDisk and Foremost for data recovery.

Another independent acquisition tool that may be used to clone disks and make forensic pictures is Guymager. Guymager, created by Guy Voncken, is entirely open source, only works on Linux-based hosts, and shares many of the same capabilities as DCLDD. Guymager, the forensic imager included in this package, is made to operate quickly, support a variety of image file types, and be extremely user-friendly. It leverages parallel compression in its high-speed multi-threaded engine to maximize performance on multi-processor and hyper-threading engines. It is shown in Fig. 5.

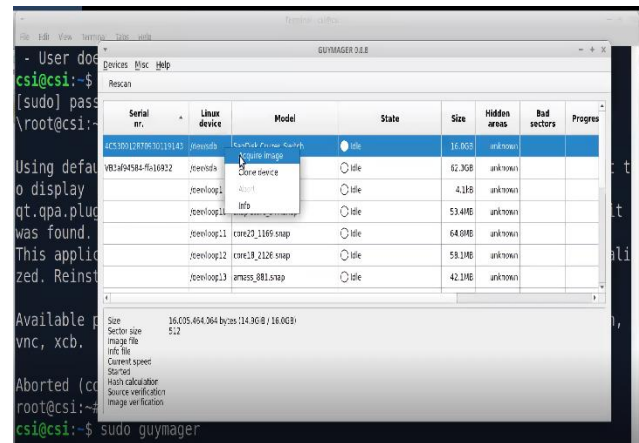


Fig. 5. The Process of Mounting File in Guymager.

1) Recovery Process

The recovery step involves a data recovery procedure using a number of programs for comparison, including TestDisk Recovery, TSK Recovery, FTK Imager, and Foremost. After that, the recovery process will use the extracted files. One of the Linux CLI-based tools for data recovery is the TSK recover utility. There are some JPG files that cannot be accessed, as illustrated in Fig. 6; however the recovery procedure utilizing the TSK recover tool was successfully recovered.

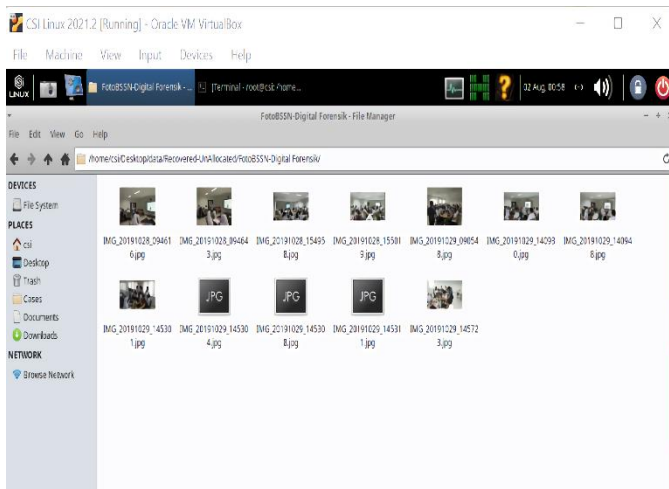


Fig. 6. File Recovery Process on TSK Recover.

A tool that is frequently used for imaging files is the FTK Imager. The FTK Imager has a number of features, including:

- Functions & Features
- Full Disk Forensic Image
- File Decryption & Password Crack
- Parsing Registry Files
- Collect, Process, and Analyze Data Sets Containing Apple's File System
- Locate, Manage and Filter Mobile Data
- Visualization Technology

This utility is frequently used to restore erased data. However, it is clear that allocated and unallocated files differ throughout the recovery step. As seen in Fig. 7, it is a file that can be recovered but cannot be opened.

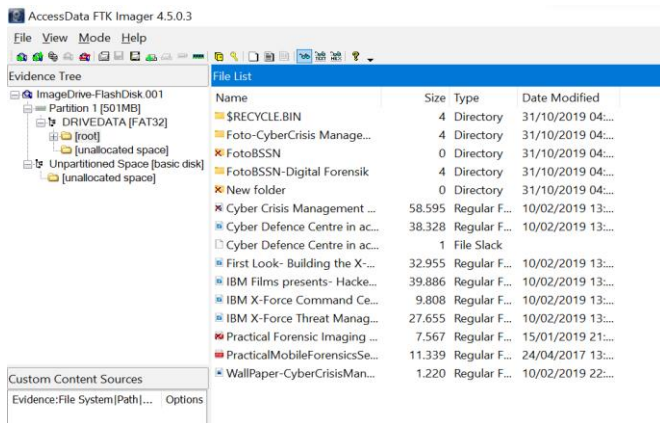


Fig. 7. File Recovery Process on FTK Imager.

The Foremost Recovery Tool is a program created to read and copy certain areas of the disk straight into the computer's memory while ignoring the underlying file system type. For

the majority of recovers, it employs a technique called file carving to search for header file types that coincide with those contained in the primary configuration file. The best recovery tool for JPG files has been used to successfully recover all of the files. It is shown in Fig. 8.

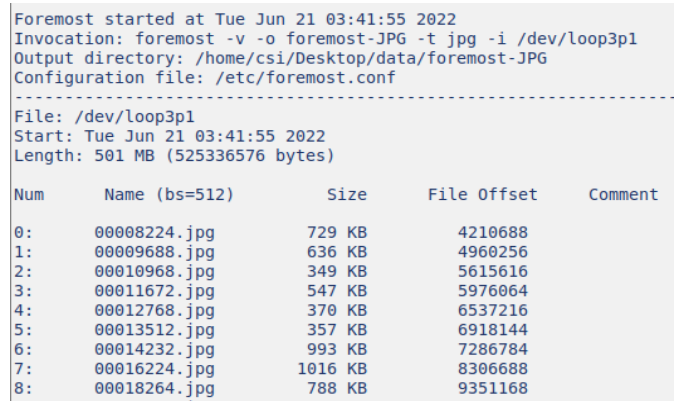


Fig. 8. File Recovery Process on Foremost Recover.

Data can be recovered from lost or deleted partitions using the free and open-source Testdisk recover utility. A digital forensics specialist can restore partitions that are unable to boot due to reasons including malware attacks and purposeful or unintentional loss of the partition table using this CLI-based application, which does not have a user interface version. Fig.9 illustrates how the recovery procedure using the Testdisk recover program was successful in restoring lost and damaged files.

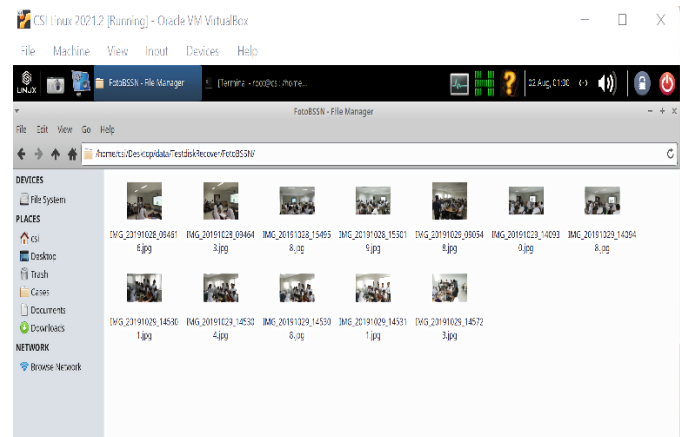


Fig. 9. File Recovery Process on Testdisk Recover.

### C. Analysis

The analysis step is where the outcomes of the files that have been checked are examined. A comparison of the data recovery tools, when situations like missing or damaged files arise, you must use the resources at your disposal to find a solution. You must experiment with all of the forensic tools, not just one. It makes sense that certain tools are unable to recover files perfectly while others are successful in doing so, as in this study is shown by Tables V and VI.

TABLE V. RESULTS OF STATUS RECOVERY ALLOCATED FILES

No	Tools	Status File Recovery		
		JPG	PNG	MP4
		Allocated		
1	TSK Recover	100%	100%	100%
2	FTK Imager	100%	100%	100%
3	Foremost Recover	100%	100%	100%
4	TestDisk	100%	100%	100%

A sort of file storage known as allocated space is still accessible, and the files contained therein can still be read logically. All files in the designated space can be fully recovered after conducting research with the aforementioned instruments.

TABLE VI. RESULTS OF UNALLOCATED FILES RECOVERY STATUS

No	Tools	Status File Recovery		
		JPG	PNG	MP4
		UnAllocated		
1	TSK Recover	50%	50%	100%
2	FTK Imager	50%	50%	100%
3	Foremost Recover	100%	100%	100%
4	TestDisk	100%	100%	100%

Files that are no longer accessible or have been erased and cannot be read logically are stored in unallocated space. Not all files have been totally and flawlessly retrieved after utilizing the following utility to do research on data files in unallocated space

#### IV. CONCLUSION

Based on the results of research on the comparison of data recovery using open source-based tools on Linux, the results of the comparison of these tools with previous research are very different. Due to the limited features available in open source forensic tools like the TSK recover tool and FTK Imager; it makes investigators hard to get valid evidence. It can be concluded that among these tools there are those that can recover data files that have been damaged and can be reopened in their entirety and some are not. One of the open source based tools that can be used is foremost recover and Testdisk recover. This tool is a solution to the problem of

recovery. Of the tools that have been tested, only 50% have been fully recovered. Namely, TSK recover and FTK imager. While the foremost tool, Testdisk, can recover 100% completely. However, tools that can't recover completely don't mean they're not good. These tools are still recommended and can be used to assist investigators in the investigation process. Investigators can have several options for forensic tools to carry out the investigative process. This study aims to determine the forensic tools that are useful today and in the future.

#### REFERENCES

- [1] P. Dibb and M. Hammoudeh, "Forensic data recovery from android os devices: An open source toolkit," Proc. - 2013 Eur. Intell. Secur. Informatics Conf. EISIC 2013, no. May, p. 226, 2013, doi: 10.1109/EISIC.2013.58.
- [2] M. Breeuwsma and M. De Jongh, "Forensic data recovery from flash memory," Small Scale Digit. ..., vol. 1, no. 1, pp. 1–17, 2007, [Online]. Available: [http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.135.5697&am p;rep=rep1&amp;type=pdf%5Chttp://www.ssddfj.org/papers/SSDDFJ\\_V1\\_1\\_Breeuwsma\\_et\\_al.pdf](http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.135.5697&am p;rep=rep1&amp;type=pdf%5Chttp://www.ssddfj.org/papers/SSDDFJ_V1_1_Breeuwsma_et_al.pdf)
- [3] Y. Guo and J. Slay, "Chapter 21 DATA RECOVERY FUNCTION TESTING," Ifip Int. Fed. Inf. Process., pp. 297–311, 2010.
- [4] J. Buchanan-Wollaston, T. Storer, and W. Glisson, "Comparison of the Data Recovery Function of Forensic Tools," IFIP Adv. Inf. Commun. Technol., vol. 410, pp. 331–347, 2013, doi: 10.1007/978-3-642-41148-9\_22.
- [5] I. P. A. E. Pratama, "Computer Forensic Using Photorec for Secure Data Recovery Between Storage Media: a Proof of Concept," Int. J. Sci. Technol. Manag., vol. 2, no. 4, pp. 1189–1196, 2021, doi: 10.46729/ijstm.v2i4.256.
- [6] M. P. Mohite and S. B. Ardhapurkar, "Design and implementation of a cloud based computer forensic tool," Proc. - 2015 5th Int. Conf. Commun. Syst. Netw. Technol. CSNT 2015, pp. 1005–1009, 2015, doi: 10.1109/CSNT.2015.180.
- [7] J. Plum and A. Dewald, "Forensic APFS file recovery," ACM Int. Conf. Proceeding Ser., 2018, doi: 10.1145/3230833.3232808.
- [8] Y. Guo, J. Slay, and J. Beckett, "Validation and verification of computer forensic software tools-Searching Function," Digit. Investig., vol. 6, no. SUPPL., pp. S12–S22, 2009, doi: 10.1016/j.diin.2009.06.015.
- [9] J. N. Hilgert, M. Lambertz, and D. Plohmann, "Extending the Sleuth Kit and its underlying model for pooled storage file system forensic analysis," DFRWS 2017 USA - Proc. 17th Annu. DFRWS USA, vol. 22, pp. S76–S85, 2017, doi: 10.1016/j.diin.2017.06.003.
- [10] I. Riadi, S. Sunardi, and S. Sahiruddin, "Analisis Forensik Recovery pada Smartphone Android Menggunakan Metode National Institute Of Justice (NIJ)," J. Rekayasa Teknol. Inf., vol. 3, no. 1, p. 87, 2019, doi: 10.30872/jurti.v3i1.2292.
- [11] M. S. Simanjuntak and J. Panjaitan, "Analisa Recovery Data Menggunakan Software," J. Tek. Inform. Komput. Univers., vol. 1, no. 1, pp. 26–32, 2021.
- [12] R. Wilson and H. Chi, "A case study for mobile device forensics tools," Proc. SouthEast Conf. ACMSE 2017, pp. 154–157, 2017, doi: 10.1145/3077286.3077564.
- [13] I. Zuhriyanto, A. Yudhana, and I. Riadi, "Analisis Perbandingan Tools Forensik pada Aplikasi Twitter Menggunakan Metode Digital Forensics Research Workshop," J. Resti, vol. 1, no. 3, pp. 829–836, 2017.
- [14] H. Handrizal, "Analisis Perbandingan Toolkit Puran File Recovery, Glary Undelete Dan Recuva Data Recovery Untuk Digital Forensik," J-SAKTI (Jurnal Sains Komput. dan Inform., vol. 1, no. 1, p. 84, 2017, doi: 10.30645/j-sakti.v1i1.31.
- [15] I. Riadi, Sunardi, and Sahiruddin, "Perbandingan Tool Forensik Data Recovery Berbasis Android Menggunakan Metode Nist," J. Teknol. Inf. dan Ilmu Komput., vol. 7, no. 1, pp. 197–204, 2020, doi: 10.25126/jtiik.202071921.

- [16] J. Panjaitan and A. C. Sitepu, "Analisis Kinerja Forensic Acquisition Tools Untuk," vol. 1, no. 2, pp. 17–25, 2021.
- [17] D. S. I. Krisnadi, "Citra Forensik Dari Barang Bukti Elektronik Dengan Metode Physical Menggunakan Acquisition Tools Tableau Imager Dan Ftk Imager," p. 16, 2020, [Online]. Available: [https://d1wqtxts1xzle7.cloudfront.net/64999902/Tableu\\_Imager\\_dan\\_FT\\_Imager.pdf?1606003446=&response-content-disposition=inline%3B+filename%3DCitra\\_Forensik\\_dari\\_barang\\_bukti\\_elektro.pdf&Expires=1609391012&Signature=ggq3RFIjWBmjsEj5dsc0ammrrNiznpH1oGNpK57](https://d1wqtxts1xzle7.cloudfront.net/64999902/Tableu_Imager_dan_FT_Imager.pdf?1606003446=&response-content-disposition=inline%3B+filename%3DCitra_Forensik_dari_barang_bukti_elektro.pdf&Expires=1609391012&Signature=ggq3RFIjWBmjsEj5dsc0ammrrNiznpH1oGNpK57)
- [18] L. M. O. Campos, E. Gomes, and H. P. Martins, "Forensic Expertise in Storage Device USB Flash Drive: Procedures and Techniques for Evidence," *IEEE Lat. Am. Trans.*, vol. 14, no. 7, pp. 3427–3433, 2016, doi: 10.1109/TLA.2016.7587651.
- [19] R. Ruuhwan, I. Riadi, and Y. Prayudi, "Penerapan Integrated Digital Forensic Investigation Framework v2 (IDFIF) pada Proses Investigasi Smartphone," *J. Edukasi dan Penelit. Inform.*, 2016, doi: 10.26418/jp.v2i1.14369.
- [20] R. Umar, I. Riadi, and G. M. Zamroni, "Mobile forensic tools evaluation for digital crime investigation," *Int. J. Adv. Sci. Eng. Inf. Technol.*, vol. 8, no. 3, pp. 949–955, 2018, doi: 10.18517/ijaseit.8.3.3591.
- [21] W. Jo, H. Chang, and T. Shon, "Digital forensic approach for file recovery in Unix systems: Research of data recovery on Unix file system," *Proc. 2016 IEEE Inf. Technol. Networking, Electron. Autom. Control Conf. ITNEC 2016*, pp. 562–565, 2016, doi: 10.1109/ITNEC.2016.7560423.
- [22] M. Riskiyadi, "Investigasi Forensik Terhadap Bukti Digital Dalam Mengungkap Cybercrime," *Cyber Secur. dan Forensik Digit.*, vol. 3, no. 2, pp. 12–21, 2020, doi: 10.14421/csecurity.2020.3.2.2144.
- [23] Anton Yudhana, Abdul Fadlil, and M. R. Setyawan, "Analysis of Skype Digital Evidence Recovery based on Android Smartphones Using the NIST Framework," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 4, no. 4, pp. 682–690, 2020, doi: 10.29207/resti.v4i4.2093.
- [24] I. A. Plianda and R. Indrayani, "Analisa dan Perbandingan Performa Tools Forensik Digital pada Smartphone Android menggunakan Instant Messaging Whatsapp," *J. Media Inform. Budidarma*, vol. 6, no. 1, p. 500, 2022, doi: 10.30865/mib.v6i1.3487.
- [25] W. Pranoto, "Penerapan Metode Live Forensics Untuk Akuisisi Pada Solid State Drive ( SSD ) NVMe Fungsi TRIM," 2020.

# Advanced Persistent Threat Attack Detection using Clustering Algorithms

Ahmed Alsanad<sup>1</sup>, Sara Altuwaijri<sup>2</sup>

Department of Information Systems  
College of Computer and Information Sciences  
King Saud University, Riyadh 11543, Saudi Arabia

**Abstract**—Advanced Persistent Threat (APT) attack has become one of the most complex attacks. It targets sensitive information. Many cybersecurity systems have been developed to detect the APT attack from network data traffic and request. However, they still need to be improved to identify this attack effectively due to its complexity and slow move. It gets access to the organizations either from an active directory or by gaining remote access, or even by targeting the Domain Name Server (DNS). Nowadays, many machine learning (ML) techniques have been implemented to detect APT attack by using the tools in the market. However, still, there are some limitations in terms of accuracy, efficiency, and effectiveness, especially the lack of labeled data to train ML methods. This paper proposes a framework to detect APT attacks using the most applicable clustering algorithms, such as the APRIORI, K-means, and Hunt's algorithm. To evaluate and compare the performance of the proposed framework, several experiments are conducted on a public dataset. The experimental results showed that the Support Vector Machine with Radial Basis Function (SVM-RBF) achieves the highest accuracy rate, reaching about 99.2%. This accurate result confirms the effectiveness of the developed framework for detecting attacks from network data traffic.

**Keywords**—APT Attack detection; DNS; network; cybersecurity; clustering algorithms

## I. INTRODUCTION

People and organizations worldwide use technology for most of their daily activities. This change is called digital transformation, which requires organizations to profoundly transform their business model, infrastructure, processes, and culture. So, the usage of the Internet is increased [1]. Although technologies and the Internet make life easier, they have been used for harmful purposes. Cybersecurity crimes impact society [2] since these crimes occur through modern communication devices using internet connections. The actors who cause a cybercrime are called attackers, and they are different kinds and have multiple goals; one of those kinds is APT Attacks. APT stands for Advanced Persistent Threat [3], and it is one of the top cybersecurity concerns in enterprise networks [4]. APT means: Advanced, which means the attacker is stealing, targeting, and data-focused attacks [5]. Persistent means an attacker identifies the target to breach, hide, and exploit them [6]. Word Threat in APT means the extraction of critical data [5]. APT are complex, and they are well-planned security attacks [7]. So, its consequences will impact the organizations by stealing intellectual property, compromising and stealing sensitive information, stealing

classified data, critical organizational infrastructures, and accessing diplomatic communication channels. Also, the ability to detect APT activity at the network level is heavily dependent on leveraging threat intelligence [8]. Attackers use multiple techniques to hide and infect the targets; the method is not limited to phishing, zero-day attack, waterhole attacks [3], and denial of service (DoS) attacks [9]. APT attack functions are developed to avoid detection as long as possible [10]. So, many techniques have been used to detect change controlling, sandboxing, and network traffic analysis [11].

Increasing the frequency of security breaches and cyberattacks on the Internet of Things (IoT) requires dependable security solutions [12]. In addition to firewalls, the Network Intrusion Detection System (NIDS) is the second network infrastructure security system that detects malicious activity and prevents attacks [13-15]. Moreover, security administrators typically choose password protection systems, encryption techniques, and access controls to protect the network. These measures, however, are insufficient to protect the system [16]. As a result, the administrators prefer to utilize Intrusion Detection Systems (IDSs) to monitor network traffic and detect malicious attacks [17-21]. For example, in [22], the authors proposed an over-sampling Principal Component Analysis (PCA) to address the anomaly detection problem.

Today, alert correlation is done using Security Information and Event Management (SIEM) systems such as Splunk, LogRhythm, and IBM QRadar [23]. They collect multiple log events and alert various sources. But the APT Attack has evolved to bypass security mechanisms that are difficult for technologies to find [24]. This paper studies how to detect APT attacks according to the framework. Currently, there is a significant potential for cyber-attack these days. A cyber-attack is intentionally exploiting computer systems, infrastructures, and networks. Cyber-attack has been done throw the attackers; the attackers are multiple kinds and category. These attackers are different from each other in terms of the goals and methods they use. Common types of cybersecurity attacks are malware attacks, Denial-of-service attacks, password attacks, and APT attacks. APT is a complex and multi-stage attack. Since its complex, they need many stages to meet their target by collecting information as much as possible and carefully [25]. Afterward, they will use their technique to reach what they need, such as phishing. Attackers then collect confidential data using multiple malware after they breach the network. Also, they use various techniques to send the data taken to another server.

Based on the NIST framework, cybersecurity programs have five primary functions. These functions are, identify, protect, detect, respond and recover. Some attackers will be known in the preserve or prevent phase, and others in the detection phase. In this paper, we focus on detecting APT attacks. That detecting APT attacks is challenging because it defeats and supersedes the premier defense devices by injecting their techniques as part of large normal traffic [25]. They are also closely linked to each other and are hidden, so it is usually too late to detect them. The attacker needs more time to efficiently distribute the attacker's activities and behaviors, with a challenging possibility to be detected. So, for APT attack detection, we use multiple techniques and tools to detect it using an attack signature, monitoring the network, and collecting network information.

Several techniques are implemented to detect the APT attack by using the tools in the market. These techniques are either using artificial intelligence (AI) or machine learning (ML) methods. However, still, there are limitations in terms of accuracy, efficiency, and effectiveness, especially the lack of labeled data to train ML methods. Significantly, the APT attacks are brutal to be detected. They usually target sensitive and critical data. An organization infected and exploited by APT attacks will harm and lose many essential assets or data. APT attacks are very complex due to their lifecycle and evolution complexity.

This paper's scope is to develop a framework and apply it as a tool to identify and detect APT attacks. Using machine learning for analyzing the attacker's behavior, the framework is implemented to help the cybersecurity specialist, especially those working in the Security operation center (SOC), to know and detect if their organization is breached and hacked by APT attackers. This framework will minimize the harm and impact that the APT attack will do. Also, it will be more accessible to cybersecurity vendors to build their detection tools. Through the proposed framework, the research contributions to the field can be summarized as follows.

- A framework to detect APT attacks is proposed to tackle the lack of labeled data using unsupervised clustering algorithms such as the APRIORI algorithm, Hunt's algorithm, and the K-means algorithm.
- The proposed framework is implemented on the CSE-CIC-IDS2018 dataset for achieving the performance of supervised learning of the ML models.
- A comparative study of the five ML classifiers is performed to detect the APT attacks.
- The framework's performance results are evaluated using several evaluation measures on the dataset.

The rest of the paper is structured as follows: Section II gives a background for the study. Section III presents the literature review of the previous work on detecting APT attacks. Section IV explains the proposed framework to identify and detect an APT attack. Section V introduces the experiments with findings and discussions. Finally, Section VI summarizes the conclusions and future research work.

## II. BACKGROUND

An APT attack's lifecycle is more complex than other kinds of attacks. A successful APT attack can be divided into multiple stages [26]. In the first stage, the attacker will define the target by determining who he will target and why he wants to target him. Next, he will select the team members and identify the required skills. Then the attacker will find the existing tools or develop new ones he/she needs. After that, the attacker will discover who has access to what he needs and what HW/SW will use. Then, the attacker will test if he/she can detect or not by deploying a miniature version of the tool, piloting a connectivity and alarm trail, check and spotting any weaknesses. Later on, he/she will launch full fledged attack on the victim's platform.

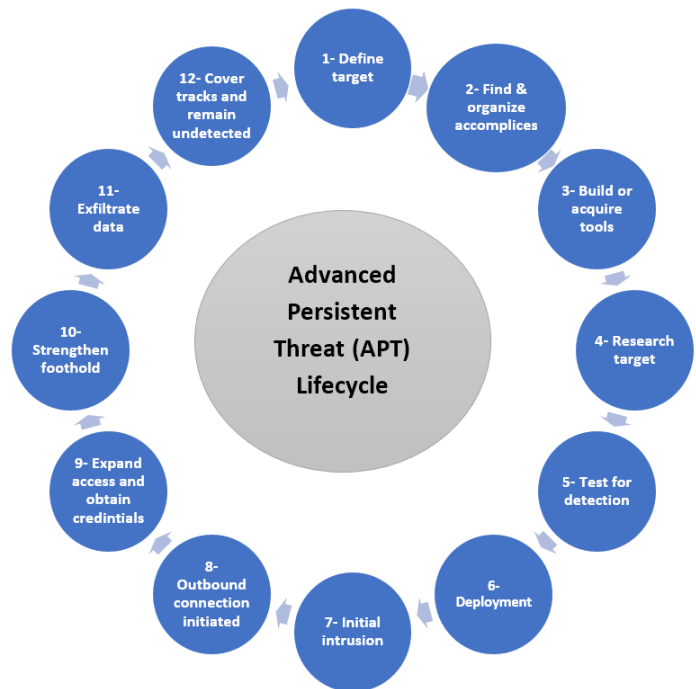


Fig. 1. APT Lifecycle.

The first entrance will be shown in the network where the target is. After that, he will establish a secure connection from victim's platform to his Command-and-Control Center. He will obtain credentials by creating a hidden Trojan on the victim platform. Then, he will start navigating through the rest of the platform to create more Trojans. After that, and once he gets what he was looking for, he will cover the tracks to remain undetected and make sure to clean up after himself. Fig. 1 shows and summarizes the stages of the APT attacker's lifecycle.

## III. LITERATURE REVIEW

This section discusses the previous work related to detecting APT attacks. An intrusion detection system (IDS) is an inevitable line of defense against cyber threats [27]. The challenge here is that IDS lacks typical evaluation methodologies to detect this attack. This section will do a literature review for detecting APT attacks and what the target is.



To detect APT attacks, two main approaches are commonly used, which are:

- Detection based on signature: It is a well-known technique based on the attack's signature [28] and low efficiency.
- Detection based on behavior: it is an enhanced technique that focuses on the attacker's signature and behavior [28], and its result is high efficiency and high processing costs.

The literature review of APT attack detection will be divided based on what the attacker targets the system, network, or domains.

#### A. APT Attack Targeting Active Directory

As Active Directory (AD) is a system that manages the organization's accounts for windows systems. This fact makes it a target for APT attacks since it is expected that domain admin privileges have been accessed by APT attacks.

To detect this attack, various research talks about it how to use machine learning and focus on the attacker's signatures and characteristics [29], and they are:

- Detection using authentication files: Using machine learning (Unsupervised) to analyze authentication files or monitor abnormal user behavior.
- Process-based detection: utilizing backlist in conjunction with signature-based detection in the log files [29], then false negatives will arise case the attackers manipulate file names of the tools because the file name is the main element on the signature algorithm.
- Detection through network traffic monitoring: an example of the methods is Golden Ticket [29] by traffic monitoring. But this feature is not implemented for windows systems.

The proposed approach was for outlier detection using Domain Controllers logs with machine learning related to processes. The advantages of this kind of method are their ability to detect AD attacks with high accuracy by abusing the command and tools that attackers are using. Also, because it uses only Domain Controller logs, it is very cost-effective. The target is detecting attacks that require control of admin privileges of the Domain Administrator.

The algorithms used are machine learning utilizing existing data without any programming effort. With this unsupervised learning, there is no need to provide correct answers [29]. So, no need to analyze the attacker's behavior.

After evaluating the methods, machine learning was the most appropriate algorithm for their way. The other one was preprocessing for machine learning which describes the necessary preprocessing for machine learning. Any logs that show a particular feature, such as logs with blank values, need to be eliminated because they can be identified with no value. When the attackers disguise their identity as an official Domain Administrator account, and the hijacked Domain Administrator account uses tools or commands, the attacker also uses false

detection. This means that Administrators use commands which are rarely used [29]. The APT Attack against AD is brutal to be detected since that attackers usually take advantage of processes and legitimate accounts [29].

#### B. APT Attack and Intrusion Detection Event

The prediction model for intrusion detection is based on events that show the probability of threat intrusion detection events through the prediction task [30]. After the analysis, it detects the attacks before or after a particular attack exists in a correlation [30]. By extracting the events of intrusions, a specific scenario is configured. When it takes place after detecting it, the next attack in the plan can be predicted by investigating at which stage of the attack scenario the intrusion detection events occur. That will result in enabling the prediction of the last threat [30].

The intrusion detection event based on the prediction model collects and pretreats intrusion detection events [30], extracts sessions and threads, creates scenarios of the attackers through correlation analysis, predicts intrusions, & expresses the analyzed results [30].

The prediction based on intrusion detection events leads to a search of an event on a scenario of the attacker when an intrusion detection event is detected [30]. When a single event occurs, other events can take place afterward. The issues that face intrusion detection events can be given as follows:

- Time required in prediction and verification of intrusion detection events: the daily average count of intrusion detection events was tremendous and incomparable with the duration of the collected data. So, it is necessary to extract successful attack events by time unit, attack type, and organization, distinguishing them from all intrusion detection events [30].
- Validity of prediction due to narrow gap in intrusion detection events: the time difference in the collected intrusion detection events verified their correlation was primarily within several seconds. Intrusion detection events in government organizations are managed by the enterprise system to monitor the database every five minutes [30]. So, the response to the events is primarily impossible, and the use of anticipated events is less.
- Intrusion detection data and intrusion detection rule: these rules are frequently added, modified, and deleted [30]. Even though those rules are changed, the sequential rules must be learned, and the rules must be applied to an independent prediction model based on intrusion detection events [30]. To do this task, a full-time employee needs to monitor and track it and be dedicated to this.
- Stability of intrusion detection system: the rule-based system used for monitoring cannot provide stability to detect the continuously changing types of attacks [30].

Based on the intrusion detection event model, prediction and verification of the events problems are not only of the time required, but the issues of cybersecurity threat prediction, such as problems in intrusion detection rules, intrusion detection

data [30], and the stability of IDS systems. In addition, the main problem is that it requires automated monitoring detection to predict different APT attacks.

### C. APT Attacks Targeting Network Infrastructure

Network infrastructure APT attacks are many. The first, called Moonlight Maze, targeted government networks [5]. The other one is called operation Aurora targeting cloud computers [5]. To detect those, the author wrote that many challenges would be faced; these challenges are [5]:

- Unsupervised anomaly-based detection approaches to discover all anomalies.
- Supervised alert correlation-based approaches decide whether some attacks are related or are a portion of a more advanced APT attack.

Evaluation and training of those approaches would require entire labeled traces of networks with widespread abnormal and intrusions behaviors [5] and need to be labeled specially for APT-correlated alerts. Also, there are several constraints in detecting an APT attack that targets network, such as:

- No one unique path in which all APT attack activities can be detected.
- Over time the APT attacks tend and adapt to use new tools and vulnerabilities.

The approaches of the IDS, including APT detection methods, consider the feature construction and selection stages as the first-time consuming step. The features can be built using machine learning and data mining methods or manually, such as association mining, sequence analysis, and frequent episode mining [5]. Some of their features categories are:

- Basic features: the essential attributes and features are collected from the connection of TCP/IP.
- Traffic features: the attributes and features that can be computed or extracted from concerning a window interval.
- Content features: the attributes and features that can be extracted from the data payload for suspicious behaviors.

The result of targeting network infrastructure should be focused on automated methods for APT attack detection. It can cover two types of use cases according to the essential infrastructure. In the first use case, the large enterprise networks are considered to have known attacks, such as GhostNet [5], Moonlight Maze, and attacks on cloud computing-based systems like Aurora Operation. Usually, the detection is based on the attack model. For the second use case, the goal network is typically used to extract sensitive information [5]. One of the achievements of this paper is the investigation and description of the existing methodologies and the detailed overview of APT detection approaches related to their infrastructure.

### D. APT Attack to Get Remote Access

The APT attack will get remote access to the target by embedding malware, installing them on the target's device,

connecting to the control server, and maintaining the control channel [31]. To maintain control of the contact, the heartbeat mechanism is also used [31]. They use HTTP, email protocol, and FTP [31] to get remote access. These protocols are standard protocols of application transport for communication to communicate between the inject sides and controls as hidden as possible to avoid security equipment inspection and audit [31]. Remote access is a perfect way for the anomaly to hide in the regular traffic since there are no variances between the communication of remote control and regular network application communication [32].

### E. APT Attacks based on Domain Name Server

One of the techniques to detect APT attacks is analyzing the domain name server (DNS). This is because DNS request constitutes only a tiny fraction of the overall traffic of the network, making it appropriate for analysis and investigation the large-scale networks [33]. Also, DNS traffic contains many significant features to recognize domain names that might be associated with malicious events. These features can be more enriched with related information [25].

The DNS feature extraction can be used to achieve an effective detection of APT attacks. There are three kinds of these features host, time, and domain. The APT Unsupervised Learning Detection (AULD) framework is proposed to detect APT attacks [25] using the DNS features. It can detect suspicious DNS domains with APT attacks based on unsupervised machine learning. The first step is to preprocess the collected DNS request; ten features have been extracted based on host, time, and domain. AULD framework can analyze many DNS log files and obtain the list of APT attacks. Also, it can extract the host, time, and domain features from the DNS log data regarding the behaviors of attackers during an APT attack detection [25].

The results have shown that the framework could detect APT activities effectively [25]. The list of suspicious domains can be detected by cybersecurity experts to define the entire APT attack detection process [25]. Also, the AULD framework can enable cybersecurity experts to analyze suspicious domains and block APT events as soon as possible [25].

### F. APT Attacks based on Accessing Unknown Domains

This section describes an architecture for detecting and monitoring APT attacks depending on access to unknown domains [28]. The architecture module of the APT attack detection and monitoring solution is shown in Fig. 3, and its methods are described as follows:

- Accuracy: APT attacks are prepared through email spam, social phishing, and email phishing [28] to reach their targets. The APT attack detection by unknown domains has a high accuracy result if one unknown domain has been detected. Others will send an alarm to users [28]. Accordingly, admins will take the appropriate action.
- Detection Time: APT attack detection can be handled in a real-time manner [28]. It is a very critical factor for preventing APT attacks at the early stages.

The algorithms for detecting an APT attack that targets these kinds are proposed as follows:

- Algorithms for detecting unknown domains: these algorithms will identify the malicious domains that are possibly suspicious domains of APT attacks [28].
- Algorithms for monitoring access to unknown domains: these algorithms will monitor suspicious activities. The unknown domains will be detected using this algorithm that will be checked by using the generation rule algorithms and simply monitoring techniques [28].

The system model used to monitor and detect the APT attacks of unknown domains is shown in Fig. 2. The model has a number of components, given as follows:

- Datacenter: the data center stores data, including weblogs, network traffic, and normalized data [28]. It gives information for monitoring and tracking network attacks. This extracted information is related to the activities and behaviors of the attackers.
- APT attacks monitoring and detection component: these components monitor and detect APT attacks using DNS logs [28]. The data center provides input for this component. This component includes the following:
  - Database: it is used to provide and store data, which is associated with the signatures of the attackers.
  - Processing component: This component implements the algorithms, methods, and techniques which are used for processing to detect APT attacks [28]. The output of this component is a set of APT attacks and suspicious domains.

- Alarm component: It is responsible for issuing warnings and alarms at different levels with evidence that the APT attacks penetrate the systems being monitored [28].

The architecture module of the APT attack detection and monitoring solution consists of the following components:

- Database: this includes:
  - Signatures of APT attacks database: it stores the signatures of all APT attacks.
  - Detecting result database: it saves the domains that are analyzed and collected by the unknown in the database [28].
  - Monitoring result database: the domains used for analysis DNS logs of unknown domains are stored in the database.

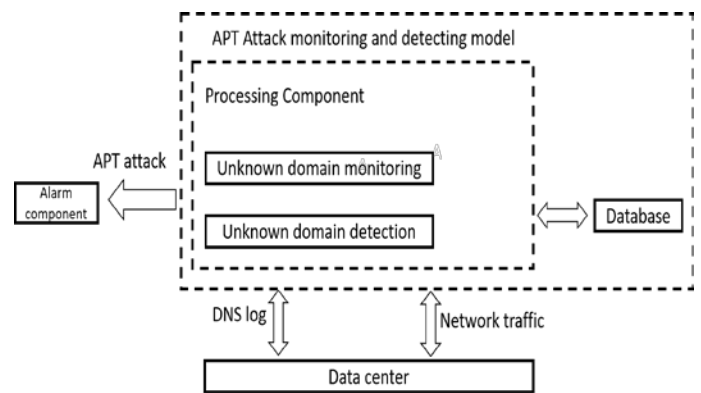


Fig. 2. Detection of APT Attacks from Unknown Domains.

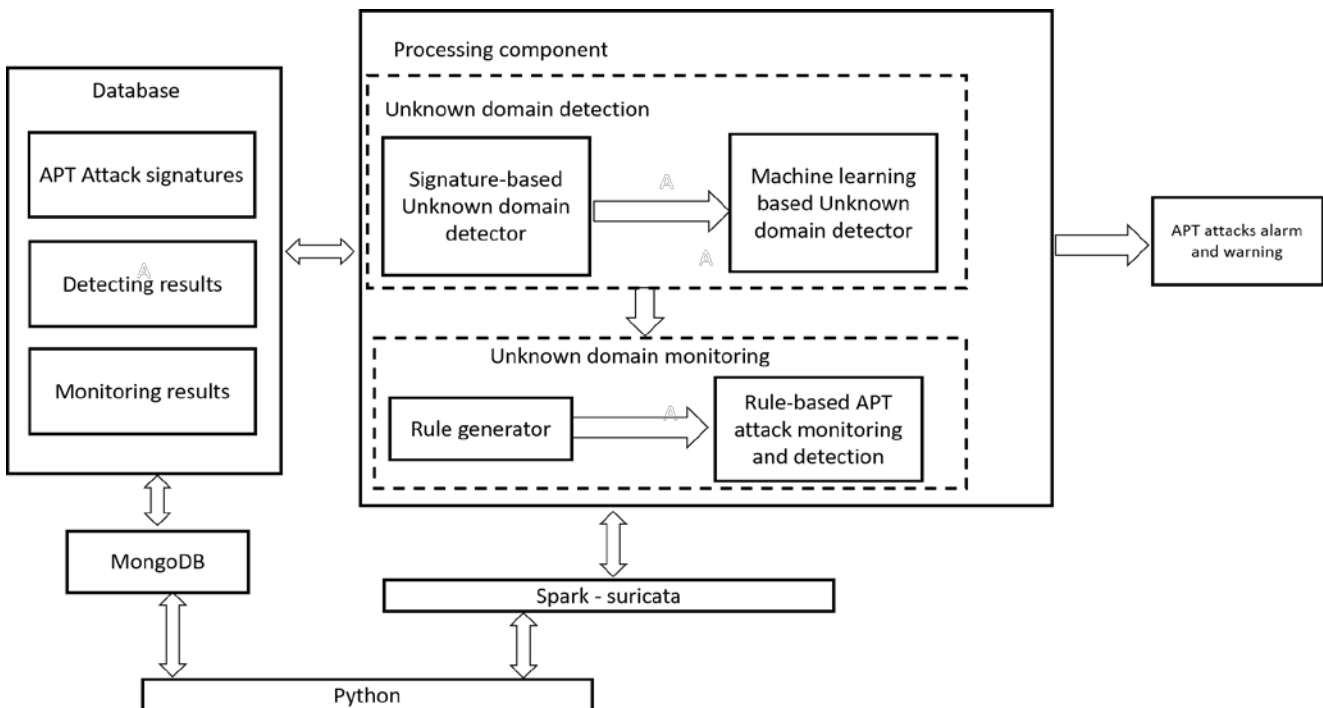


Fig. 3. Architecture Module of the APT Attack Detection and Monitoring Solution [28].

- Unknown domain detection: it consist of a set of algorithms to detect and monitor unknown domains, and it includes:
  - Signature-based unknown domain detector: they collect and extract the signatures from the described APT attacks DNS logs. They are used as evidence for APT attack detection. It compares the domains signatures in the DNS logs with the collected actual APT attack signatures. If the signatures are matched, then these domains are malicious; otherwise, they are benign [28].
  - Machine learning-based unknown domain detector: this is done to identify unknown APT attack domains. A set of suspicious domains is provided as input, and a set of unknown malicious/benign domains is returned [28]. In this study, a clustering technique was employed.

As a result, the APT attack has multiple stages and steps of its implementation. If one stage fails, the whole APT will fail [28]. The method presented is for APT attacks that uses monitoring access to the unknown domains in a real-time manner in high efficiency and effectiveness.

The persistent nature of this kind of attack reveals the necessity of having precise analysis to measure the damages in the absence of proper diagnosis and treatment. This raises several concerns:

- 1) Continues activities from adversaries to breach victims' platforms and to seek the weakest link. This requires continuous monitoring activities and applying the rights update to the media.
- 2) It is not easy to detect the breaches once the advisories gain access to the victims' platform. This requires specialized tools and skilled human resources.
- 3) Recovery will take time to clean up all the resources because of the methods used during the breach.
- 4) Cost again this type of attack is high since it requires advanced detection and protection tools and continuous monitoring.
- 5) Skilled resource availability will be playing a significant role, and it has to be appropriately addressed.

#### IV. PROPOSED FRAMEWORK TO IDENTIFY AND DETECT AN APT ATTACK

APT attacks are complex and hard to be detected. This paper introduces a framework for identifying and detecting APT attacks. The framework is an automated unsupervised machine learning [25], and the output is a set of suspicious DNS domains by analyzing the DNS features. This framework can report the suspicious domains to the security engineer and help the defenders detect faster APT attacks [25]. The framework is divided into four stages: the data collection stage, data preprocessing stage, feature extraction stage, and clustering stage. Fig. 4 shows the flowchart of the proposed framework.

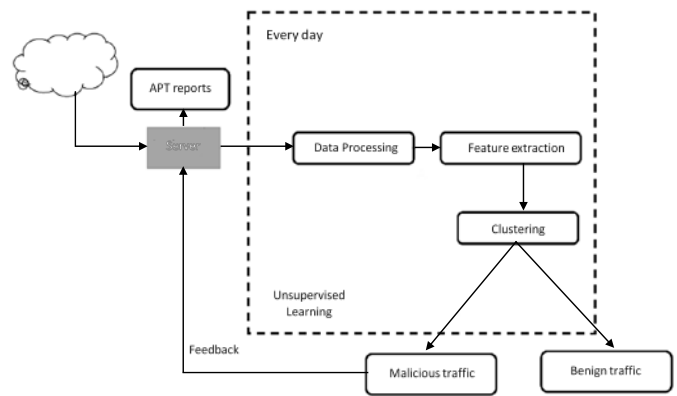


Fig. 4. Flowchart Diagram of the Proposed Framework.

The first part of this framework is data collection, which will collect DNS data log records for a certain period. When finding a precise time sequence for an IP of the internal host, the accessing date, the accessing domain, and other fields among the APT attacker's reports and giving some malicious domains, this will star detect the APT attack.

The second part is data processing; in this part, we will do the following [25]:

- By extracting a valid field and changing the format of the data in the data raw.
- Folding domain into the next level of domain.
- Deleting the whitelist of sites.
- Deleting famous websites within the internal network to get the experimental data.

A feature extraction will then be done by knowing the number of devices that get access to the domains, the domain's popularity, access time, automatic connection, domain age, and similarity of a domain. This is all based on the three types of features, which are time, host, and domain. The last part is the clustering process, and this is done according to the proper algorithm upon testing them such as K-mean clustering algorithm, or Hierarchical clustering, or Density-based clustering algorithms. The framework contains the following steps:

- Data preparation: This is the stage of data preprocessing in which unnecessary features and duplicate instances are removed in preparation for identification. Convert categorical attributes to numerical values through data digitization. Normalization for modifying the scale, type, and probability distribution of variables in a dataset is an example of a data transformation.
- Feature selection and reduction: using the PCA technique to pick the most relevant features subset approaches the detection phase as input.

Detection: On the CSE-CIC-IDS2018 dataset, we improved classification accuracy by utilizing KNN, decision tree, and two kernels (linear, RBF) with SVM machine learning classifiers, as well as a random forest classifier.

### A. Model Evaluation

After the model is trained using the training data samples, it can pass into the test step. Inspecting how the model works in practical circumstances is the aim of testing. This stage allows us to evaluate the model's precision. In this study, the model attempts to identify the APT attack using the knowledge gained during the training step. The evaluation process is vital because it enables us to determine whether the model accomplishes the objective of classifying the network traffic. The previous procedures must be repeated until the requisite accuracy is attained if the model does not function as anticipated during the testing phase. As previously indicated, it should not use the same data that was used during the training phase. It needs to utilize a different data splitter from the data set for analysis.

The accuracy of the outcome is one of the classification measures taken into consideration for evaluating the trained models. When producing classification outputs, there are four possible outcomes: true positives, true negatives, false positives, and false negatives. These four outcomes are represented on a confusion matrix. The matrix can be created based on the results after classifying the test inputs, and each output can be classified as one of the potential outcomes. The model's accuracy is measured by the proportion of correct classification from the test data. The number of correctly classified instances divided by the total number of instances gives the result of accuracy. Additionally, classification models are assessed using additional metrics such as precision and recall.

### B. Adopted Algorithms in the Framework

This section proposes an intrusion detection system based on machine learning algorithms. A Principal Component Analysis (PCA) algorithm is used for feature reduction. This method improves the performance detection task [22]. Traditionally, PCA reduces the feature dimension by linearly transforming original  $n$ -dimensional features into  $n$  orthogonal axis, as shown in Fig. 5. By projecting an observation onto each of these axes, a new set of  $n$  uncorrelated variables is created. The new feature vector is composed of a subset of these variables with a high eigenvalue. However, each derived feature requires  $n \times n$  multiplications and the use of all original features to compute. The computation time for feature extraction will rise as a result of this. In this study, a PCA removes some unnecessary features from the feature set. The information extracted from the coefficients of the Principal Components (PC) is used for feature ranking and reduction.

The covariance matrix ( $C$ ), of the  $n$ -dimensional features vector taken from positive training samples, is created first. The Principal Components are then determined using  $C$ 's eigenvalues and eigenvectors (PCs). There are a total of  $n$  potential PCs. Each of the PCs has  $n$  coefficients, each of which is associated with a correlated feature from the original feature pool. The characteristic associated with the PC's largest coefficient is placed in the highest rank by starting with the first PC. The same technique is used on succeeding PCs to generate a list of features in descending order. A varying number of low-ranked features are deleted depending on the ranking to generate a subset of reduced features.

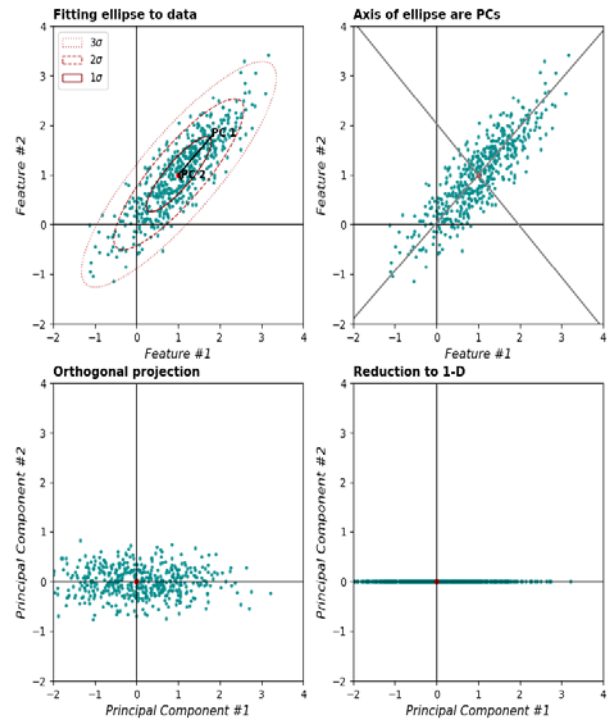


Fig. 5. PCA Feature Reduction by Linearly Transforming Original  $n$ -dimensional Features into  $n$  Orthogonal axis.

- Experiments have been conducted to determine the smallest number of characteristics that can accurately represent the entire feature set. After that, we achieved a comparative study of the five proposed classifiers, which are:
- Decision Tree (DT).
- Random Forest (RF).
- K-Nearest Neighbor (KNN).
- Support Vector Machine with Linear Function (SVM-Linear).
- Support Vector Machine with Radial Basis Function (SVM-RBF).

## V. EXPERIMENTS AND DISCUSSIONS

### A. CSE-CIC-IDS2018 Dataset

In this section, we describe the CSE-CIC-IDS2018 dataset [1] used to evaluate the proposed framework. It includes detail on intrusions as well as protocol specifics. The Canadian Institute for Cybersecurity released its most recent dataset in 2018-2019. This dataset contains seven different forms of assaults: Botnet, infiltration, DoS, Heartbleed, DDoS, Brute force, and Web attacks. The compromised firms had 30 servers and 420 PCs, while the attacking infrastructure had 50 terminals.

The CICFlowMeter-V3 dataset [26] is collected traffic of AWS network and machine log files with more than 70 extracted features. The best way to test and evaluate the system

framework is represented by the network's applications and the lowest level entities; it also refers to the move from static data to dynamic data, which is real-time traffic on the Amazon platform (AWS). Furthermore, the dataset was improved by taking into account the standards that were designed to produce CIC-IDS2017. In addition to the basic criteria, it has the following advantages:

- There are very few duplicate data records.
- Uncertain data is almost non-existent.
- The dataset is in CSV format so that it can be used immediately without further processing.

### B. Evaluation Metrics

Some evaluation metrics such as confusion matrix, accuracy, detection rate, precision, recall, and F1-score are used to evaluate the effectiveness of the framework's ML algorithms.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

$$Recall (True Positive Rate (TPR)) = \frac{TP}{TP+FN} \quad (3)$$

$$F1-Score = 2 * \frac{Precision*Recall}{Precision+Recall} \quad (4)$$

- Confusion matrix: In intrusion detection, a confusion matrix is a useful tool for predicting the type of network attack. It contains where TP refers to true positive instances (TP), true negative instances (TN), false positive instances (FP), and false negative instances (FN).
- Accuracy: The percentage of positive data cases detected correctly.
- Precision: The number of attacks correctly returned.
- Recall or True Positive Rate (TPR): The number of attacks the system returns.
- F1-score: In our approach, the rate of precision and recall:

### C. Experimental Results

The CSE-CIC-IDS2018 dataset [1] is first preprocessed by eliminating eleven non-essential features such as the timestamp, average number of bulk rates, and number of times the PSH flag was set in packets. The parameters are set by default in all of the implemented algorithms in this report, with the exception of KNN, which uses the n nearest neighbor's property ( $n = 3$ ). The number of classes in the suggested algorithm was determined to be one (zero for non-attack types and one for attack types). The most recent dataset available was used for training and testing the CSE-CIC-IDS2018 dataset. In the trials, training and test data were divided into 80 percent and 20 percent to assess the performance results related to training and testing.

Although PCA aims to maximize the distance between data points, it has no concept of classes. The default libraries in

Python programming language like the Scikit-Learn library, are used. In the experiments, most of the hyper-parameters for machine learning algorithms were set to default. Table I shows the hyper-parameter values for ML algorithms classifiers.

The accuracy definition is crucial since accuracy is an essential criterion for evaluating the efficiency of prediction systems. Accuracy is frequently used to refer to a system's perfect accuracy. However, accuracy can also relate to a class individual accuracy. For researchers working with unbalanced datasets, the definition of accuracy is the average of the accuracies of all classes, which is crucial. In this report, we used K-Nearest Neighbor (KNN) [34], Random Forest (RF) [3], linear support vector machine (SVM-linear) [31], Decision Tree (DT) [30], and Radial basis function (RBF) support vector machine (SVM-RBF) [11] classifiers to classify and detect benchmark CSE-CIC-IDS2018 intrusion detection dataset.

An intrusion detection system should ideally have a 100 percent attack % true-positive rate (TPR) and a 0% false-positive rate (FPR). However, it is difficult to achieve in practice. Table II and Fig. 6 depict the results of these metrics.

The SVM-RBF classification algorithm, as shown in Table II, is the most successful, with a 99.2% accuracy rate. With a 99.1% accuracy rate, the RF classifier algorithm is the second most efficient. Finally, the DT classifier, which had the lowest accuracy rate of 94.2% was applied to the proposed dataset.

With a precision rate of 99.9%, the random forest classifier classification algorithm, as indicated in Table II, is the most successful. The SVM-RBF algorithm is the second most efficient, with a 99.3 % precision rate. Finally, when applied to the proposed dataset, the DT classifier had the lowest precision rate of 79.9%.

TABLE I. MACHINE LEARNING CLASSIFIERS HYPER-PARAMETER VALUES

Algorithm	Hyper-parameter
Decision Tree (DT)	criterion='gini', splitter='best', min_samples_split=2
Random Forest (RF)	n_estimators=1000, criterion='gini', min_samples_split=2, min_samples_leaf=1
K-Nearest Neighbor (KNN)	n_neighbors=3, weights='uniform', leaf_size=30, metric='minkowski'
Support Vector Machine (SVM-linear)	Regularization parameter (C) =1, kernel='linear'
Support Vector Machine (SVM-RBF)	Regularization parameter (C) =1, kernel='rbf'

TABLE II. COMPARISON OF THE RESULTS USING FIVE MACHINE LEARNING CLASSIFIERS

ML algorithm	Accuracy	Precision	F1-score	TPR
DT	0.942	0.799	0.865	0.941
RF	0.991	0.999	0.976	1.000
KNN	0.970	0.888	0.926	0.970
SVM-Linear	0.960	0.853	0.905	0.959
SVM-RBF	0.992	0.993	0.979	0.998



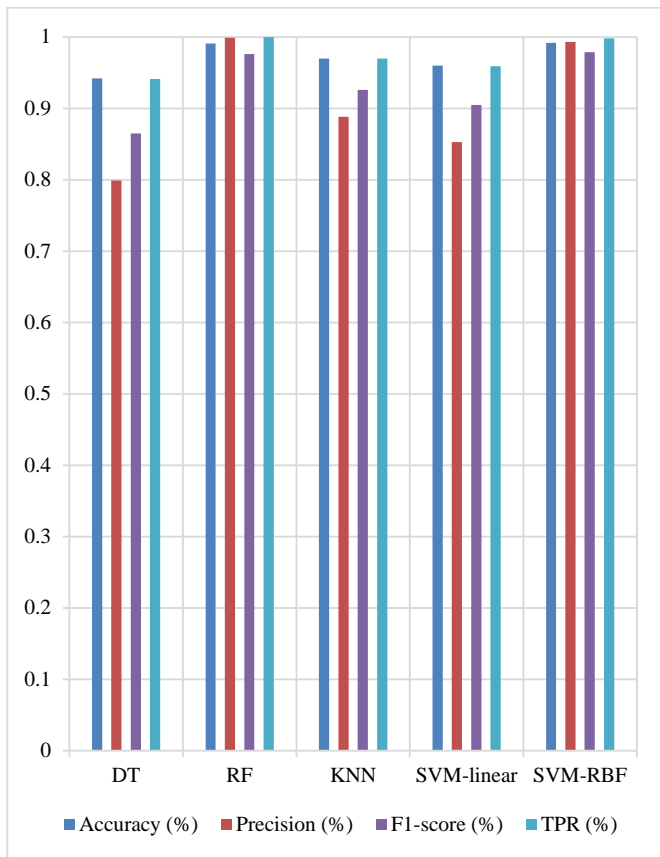


Fig. 6. Performance Analysis of Proposed Framework.

The KNN classifier classification algorithm, as seen in Table II, has the highest recall rate of 96.8%. The second most efficient approach is the SVM-RBF, which has a 96.6% Recall rate. Finally, the DT classifier had the lowest recall rate of 94.3%.

With a 97.9% F1-score rate, the SVM-RBF classification technique, as indicated in Table II, is the most successful. The RF classification algorithm is the second most efficient, with an F1-score rate of 97.6%. Finally, when applied to the provided dataset, the DT classifier had the lowest F1-score rate of 86.5%.

The RF classification algorithm, as shown in Table II, it is the most successful, with a true-positive rate (TPR) of 100%. With a TPR rate of 99.8%, the SVM-RBF algorithm is the second most efficient. Finally, the DT classifier had the lowest TPR rate of 94.1% when applied to the proposed dataset.

## VI. CONCLUSIONS AND FUTURE WORK

The APT attack is not easy or soft kind of attacker. So, detecting it in the early stages will reduce the organization's impact after exploiting it. Also, detecting it using the security exiting tools throw the proposed framework will let it done in a systematic approach. Because of the widespread usage of the Internet in recent years, computational devices can now connect to the universal network from anywhere. However, the anonymous nature of the Internet leads to numerous security flaws in the network, resulting in intrusions. Modern attackers are more intelligent, and they may create new malware and

malicious code with the assistance of automated development tools, depending on the limited capability of IDS. This paper uses data transformation and normalization with a reduction procedure using PCA. The benchmark CSE-CIC-IDS2018 dataset is consisted of five different machine learning classifiers for malware IDS detection (DT, RF, KNN, SVM-Linear, and SVM-RBF). The experimental finding showed that the proposed models had a satisfactory performance, specifically when using Random Forest and support vector machine with Radial basis function classifiers, which have a 100% true-positive rate. Several machine learning methods are being transferred to deep learning models due to the convenience of big data technologies. This paper is a preliminary experiment to see how machine learning algorithms can simply and effectively detect attacks from network data traffic. As a result, in the future, deep learning algorithms are recommended to be applied for big DNS data requests.

## ACKNOWLEDGMENT

“The author is thankful to the Deanship of Scientific Research, College of Computer and Information Sciences (CCIS) at King Saud University for funding this research.”

## REFERENCES

- [1] P. N. Bahrami, A. Dehghantanha, T. Dargahi, R. M. Parizi, K.-K. R. Choo et al., "Cyber kill chain-based taxonomy of advanced persistent threat actors: Analogy of tactics, techniques, and procedures," *Journal of information processing systems*, vol. 15, no. 4, pp. 865-889, 2019.
- [2] A. S. Ahmed, S. Deb, A.-Z. S. B. Habib, M. N. Mollah and A. S. Ahmad, "Simplistic approach to detect cybercrimes and deter cyber criminals," in *2018 International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2)*, 2018, pp. 1-4: IEEE.
- [3] S. Li, Q. Zhang, X. Wu, W. Han and Z. Tian, "Attribution classification method of apt malware in iot using machine learning techniques," *Security Communication Networks*, vol. 2021, 2021.
- [4] Q. Zou, X. Sun, P. Liu and A. Singhal, "An approach for detection of advanced persistent threat attacks," *Computer Communications*, vol. 53, no. 12, pp. 92-96, 2020.
- [5] B. Stojanović, K. Hofer-Schmitz and U. Kleb, "Apt datasets and attack modeling for automated detection methods: A review," *Computers Security*, vol. 92, p. 101734, 2020.
- [6] C. Do Xuan, M. H. Dao and H. D. Nguyen, "Apt attack detection based on flow network analysis techniques using deep learning," *Journal of Intelligent Fuzzy Systems*, vol. 39, no. 3, pp. 4785-4801, 2020.
- [7] Y.-x. Xie, L.-x. Ji, L.-s. Li, Z. Guo and T. Baker, "An adaptive defense mechanism to prevent advanced persistent threats," *Connection Science*, vol. 33, no. 2, pp. 359-379, 2021.
- [8] G. Zhao, K. Xu, L. Xu and B. Wu, "Detecting apt malware infections based on malicious dns and traffic analysis," *IEEE access*, vol. 3, pp. 1132-1142, 2015.
- [9] A. L. G. Rios, Z. Li, K. Bekshentayeva and L. Trajković, "Detection of denial of service attacks in communication networks," in *2020 IEEE international symposium on circuits and systems (ISCAS)*, 2020, pp. 1-5: IEEE.
- [10] T. Bodström and T. Hämmäläinen, "A novel deep learning stack for apt detection," *Applied Sciences*, vol. 9, no. 6, p. 1055, 2019.
- [11] K. A. A. Alminshid and M. N. Omar, "A framework of apt detection based on packets analysis and host destination," *Iraqi Journal of Science*, pp. 215-223, 2020.
- [12] K. Gopalakrishnan, "Security vulnerabilities and issues of traditional wireless sensors networks in iot," in *Principles of internet of things (iot) ecosystem: Insight paradigm*: Springer, 2020, pp. 519-549.

- [13] S. Mishra, R. Sagban, A. Yakoob and N. Gandhi, "Swarm intelligence in anomaly detection systems: An overview," *International Journal of Computers Applications*, vol. 43, no. 2, pp. 109-118, 2021.
- [14] B. B. Zarpelão, R. S. Miani, C. T. Kawakani and S. C. de Alvarenga, "A survey of intrusion detection in internet of things," *Journal of Network Computer Applications*, vol. 84, pp. 25-37, 2017.
- [15] I. Sharafaldin, A. H. Lashkari and A. A. Ghorbani, "Toward generating a new intrusion detection dataset and intrusion traffic characterization," *ICISSp*, vol. 1, pp. 108-116, 2018.
- [16] H. Tabrizchi and M. J. T. j. o. s. Kuchaki Rafsanjani, "A survey on security challenges in cloud computing: Issues, threats, and solutions," vol. 76, no. 12, pp. 9493-9532, 2020.
- [17] B. Reis, S. B. Kaya, G. Karatas and O. K. Sahingoz, "Intrusion detection systems with gpu-accelerated deep neural networks and effect of the depth," in *2018 6th International Conference on Control Engineering & Information Technology (CEIT)*, 2018, pp. 1-8: IEEE.
- [18] G. Karatas, O. Demir and O. K. Sahingoz, "Increasing the performance of machine learning-based idss on an imbalanced and up-to-date dataset," *IEEE Access*, vol. 8, pp. 32150-32162, 2020.
- [19] V. Kanimozhi and T. P. Jacob, "Calibration of various optimized machine learning classifiers in network intrusion detection system on the realistic cyber dataset cse-cic-ids2018 using cloud computing," *International Journal of Engineering Applied Sciences Technology*, vol. 4, no. 6, pp. 2455-2143, 2019.
- [20] N. F. Haq, A. R. Onik, M. A. K. Hridoy, M. Rafni, F. M. Shah et al., "Application of machine learning approaches in intrusion detection system: A survey," *IJARAI-International Journal of Advanced Research in Artificial Intelligence*, vol. 4, no. 3, pp. 9-18, 2015.
- [21] Q. R. S. Fitni and K. Ramli, "Implementation of ensemble learning and feature selection for performance improvements in anomaly-based intrusion detection systems," in *2020 IEEE International Conference on Industry 4.0, Artificial Intelligence, and Communications Technology (IAICT)*, 2020, pp. 118-124: IEEE.
- [22] Y.-J. Lee, Y.-R. Yeh and Y.-C. F. Wang, "Anomaly detection via online oversampling principal component analysis," *IEEE transactions on knowledge data engineering*, vol. 25, no. 7, pp. 1460-1470, 2012.
- [23] S. M. Milajerdi, R. Gjomemo, B. Eshete, R. Sekar and V. Venkatakrishnan, "Holmes: Real-time apt detection through correlation of suspicious information flows," in *2019 IEEE Symposium on Security and Privacy (SP)*, 2019, pp. 1137-1152: IEEE.
- [24] C.-H. Liu and W.-H. Chen, "The study of using big data analysis to detecting apt attack [j]," *Journal of Computer Science*, vol. 30, no. 1, pp. 206-222, 2019.
- [25] F. J. Abdullayeva, "Advanced persistent threat attack detection method in cloud computing based on autoencoder and softmax regression algorithm," *Array*, vol. 10, p. 100067, 2021.
- [26] B. I. Messaoud, K. Guennoun, M. Wahbi and M. Sadik, "Advanced persistent threat: New analysis driven by life cycle phases and their challenges," in *2016 International conference on advanced communication systems and information security (ACOSIS)*, 2016, pp. 1-6: IEEE.
- [27] A. A. Ahmed and Y. W. Kit, "Collecting and analyzing digital proof material to detect cybercrimes," in *2018 IEEE 16th Intl Conf on Dependable, Autonomic and Secure Computing, 16th Intl Conf on Pervasive Intelligence and Computing, 4th Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress (DASC/PiCom/DataCom/CyberSciTech)*, 2018, pp. 742-747: IEEE.
- [28] D. X. Cho and H. H. Nam, "A method of monitoring and detecting apt attacks based on unknown domains," *Procedia Computer Science*, vol. 150, pp. 316-323, 2019.
- [29] W. Matsuda, M. Fujimoto and T. Mitsunaga, "Detecting apt attacks against active directory using machine leaning," in *2018 IEEE Conference on Application, Information and Network Security (AINS)*, 2018, pp. 60-65: IEEE.
- [30] Y.-H. Kim and W. H. Park, "A study on cyber threat prediction based on intrusion detection event for apt attack detection," *Multimedia tools applications*, vol. 71, no. 2, pp. 685-698, 2014.
- [31] M. Li, W. Huang, Y. Wang, W. Fan and J. Li, "The study of apt attack stage model," in *2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS)*, 2016, pp. 1-5: IEEE.
- [32] A. Ajibola, I. Ujata, O. Adelaiye and N. A. Rahman, "Mitigating advanced persistent threats: A comparative evaluation review," *Int'l J. Info. Sec. Cybercrime*, vol. 8, p. 9, 2019.
- [33] Y. Zhauniarovich, I. Khalil, T. Yu and M. Dacier, "A survey on malicious domains detection through dns data analysis," *ACM Computing Surveys*, vol. 51, no. 4, pp. 1-36, 2018.
- [34] A. Ajibola, I. Ujata, O. Adelaiye, N. A. Rahman and Cybercrime, "Mitigating advanced persistent threats: A comparative evaluation review," *Int'l J. Info. Sec. Cybercrime*, vol. 8, p. 9, 2019.

# Energy Efficient Node Deployment Technique for Heterogeneous Wireless Sensor Network based Object Detection

Jayashree Dev<sup>1</sup>, Jibitesh Mishra<sup>2</sup>

Department of Information Technology, Odisha University of Technology & Research, Bhubaneswar, India<sup>1</sup>  
Department of Computer Science & Application, Odisha University of Technology & Research, Bhubaneswar, India<sup>2</sup>

**Abstract**—Lifetime of the network and the quality of operation are the two important issues in a wireless sensor network system meant for object detection and tracking application. At the same time, there should be a tradeoff between network cost and the quality of operation as high cost of the network limits its real-time usability. Heterogeneous wireless sensor networks promises the prolonged network lifetime as well as enhances network reliability as they contain a mixture of nodes with different characteristics. Further prolongation of network lifetime can be achieved by managing the available node energy in a proper way, i.e, by minimizing number of communication, minimizing node density, minimizing overhead information generated during operation etc. Proper node deployment scheme not only helps to enhance the lifetime of the network but also helps in reducing deployment cost while maintaining the quality of operation in terms of object detection accuracy. This paper focuses on the energy efficient node deployment in heterogeneous wireless sensor network system with the features of maximum network coverage, optimum node density and optimum network cost. This paper proposes a novel energy efficient node deployment algorithm that determines the number of static and mobile nodes required for deployment and then relocates the mobile nodes to cover up the coverage hole using 8-neighbourhood and Particle Swarm Optimization (PSO) algorithm. The performance of the proposed algorithm is compared with corresponding model of Harmony Search Algorithm (HSA) and PSO based node deployment and it is seen that the proposed model outperforms better in comparison to them.

**Keywords**—Heterogeneous wireless sensor network; energy efficiency; node deployment; object detection network; particle swarm optimization; harmony search algorithm

## I. INTRODUCTION

Wireless Sensor Network (WSN) is a network of tiny connected sensors deployed in different fields like surveillance system, disaster management, wildlife monitoring, and health care system for surrounding environment information collection and processing and to initiate action according to the result of processing. According to the type of node, there are two types of sensor network:-homogeneous network and heterogeneous network. In homogeneous network, all nodes are with same characteristics whereas in case of heterogeneous network they can be of different characteristics. Heterogeneous wireless sensor networks better in comparison to homogeneous network as they support high

lifetime and high coverage.

The main constraint of WSN is limited energy. WSN meant for object detection and tracking must have sufficient lifetime to complete the desired operation and must have full coverage of the monitoring area. Otherwise, the quality of operation cannot be relied. At the same time, while designing the network for this kind operation the cost factor cannot be ignored. Having powerful nodes in a network to have large network lifetime is not sufficient to have actually the long network life .Also, the power should not be wasted in unnecessary communications and should be properly utilized. One of the way to minimize the unnecessary power usage is to have a proper node deployment scheme for the designated application. There are two important factors relating to any node deployment scheme:-node density and node location. Node density is defined as the minimum number of nodes required to cover a given area. Determining optimum node density is a NP complete problem. Low node density creates coverage hole which creates problem in achieving accuracy in result and may initiate wrong action in response to the result. Similarly, high node density increases number of routes to base station (BS) and number of communication which leads to wastage of energy. High node density also increases the system cost. Therefore, determination of optimum node density is very much important to maintain the quality of the network.

There are two types of node deployment: random deployment and deterministic deployment [1, 2]. In random deployment, sensor nodes are randomly placed in the target area and hence there is the chance of creation of coverage hole in the network. This type of deployment is well suited for large WSN and particularly in the area which is inaccessible easily to the human being. In deterministic deployment, sensor nodes are deployed in pre-calculated position and hence suitable for small size WSN. From the point of network cost factor, deterministic deployment is costly if the size of the network is kept constant for both types of deployments.

There are two types of sensing models: binary disk sensing model and probabilistic sensing model [1, 2].

There are three different types of coverage of target area:-blanket or full coverage, barrier coverage and point coverage [1, 2]. In full coverage, entire monitoring area is covered by sensors. In barrier coverage, barrier of the sensor nodes are monitored whereas in point coverage method, the point of

interest is covered by sensor nodes. Blanket coverage is most suited for object detection and tracking operation. The factors that affect the coverage are characteristic of sensor nodes and coverage algorithm adopted for the application.

This paper considers the problem of node deployment in terms of node density, maximization of coverage area, network lifetime and network cost. Node density has the direct impact on lifetime of the network, network coverage and cost of the network. For a target tracking WSN, the quality of the tracking depends on both network coverage and network lifetime. The cost of the network is directly proportional to the node density. Network coverage depends on node density and location of nodes in the monitoring area.

The rest of the paper is organized as: Section II gives a description of the related works, Section III describes the terminologies defined, problem statement, Section IV gives a description of proposed model and Section V gives the description of performance evaluation of the proposed model.

## II. RELATED WORK

Balancing the deployment quality and deployment cost is a challenging task in random deployment in wireless sensor networks. Connectivity in the network depends on the node density and network coverage. Though a lot of research works are done in past to address this issue, still no work is robust. While some papers focuses on maximization of network coverage, some focuses on cost of the deployment. But the WSN designed for object detection and tracking operation must have a balanced feature of maximum network coverage, cost of deployment and energy efficiency. This section gives a description of earlier attempts made by researchers to solve this issue.

B. A. Fuhaidi et al. [3] have proposed a node deployment model based on harmony search algorithm (HSA) and probabilistic sensing model (PSM) in which attempt is made to maximizing network coverage with minimum cost. The cost of the deployment is controlled by controlling the number of mobile nodes to be deployed in the area of interest (AOI). The area of interest is divided into a number of equal sized cells and the centre of the cell is considered as target point. Thus, AOI contain a set of target point. Static nodes are deployed using random deployment scheme initially. PSM is used to calculate the network coverage using target point set. If two or more sensors cover the same target point, then, it is said that sensors are overlapped. The probability of coverage overlapping is determined on the basis of coverage threshold. Next, mobile nodes are added according to the requirement. HSA is used to optimize the mobile node location so that overlapped region can be minimized. The authors claim that the proposed model contains less number of nodes and has maximum network coverage in comparison to homogeneous deployment and HEWSN model. No doubt, heterogeneous wireless sensor networks are better in comparison to homogeneous wireless sensor network and are more useful for the application like object detection and tracking. This feature cannot be ignored in the name of cost because quality of operation also matters. So, there is a need of balancing between different types of nodes used in the network so that quality will not be compromised. The authors have not

considered this matter. C. Zygowski and A. Jaekel [4] have proposed an algorithm based on mixed integer linear programming for effective path planning for mobile nodes to fill the coverage hole and to maximize area coverage. The algorithm focuses on minimum distance travel for mobile nodes and in minimum time. The deployment cost is not taken into consideration here. I. Alablani and M. Alenazi [5] proposed a node deployment strategy named Evaluated Delaunay Triangulation-based Deployment for Smart Cities (EDTD-SC) that focuses on sensor distribution and sink placement in smart cities. The algorithm utilizes Delaunay triangulation and k-means clustering to optimize the node location to improve coverage while maintaining connectivity and robustness with obstacles existence in the area of interest. The deployment scheme outperforms random and regular deployment in terms of network coverage. The work is suitable for small sized network and do not explore the feature of heterogeneity in node deployment. P. Prabhakaran et al.[6] have proposed an adaptive virtual force algorithm for node deployment in hybrid wireless sensor network meant for object tracking. Initially, static nodes are deployed and coverage hole is determined. Then, for hole patching mobile nodes are used. The optimum location of the mobile nodes is calculated using adaptive virtual force algorithm. The scheme does not focus on the cost of the network. J. Mao et al.[7] have proposed a partitionable polyhedral node deployment scheme for warehouse monitoring systems. This scheme proposes a node deployment collaborative perception model based on 0-1 perception model and exponential model. The 3D space is divided into a number of voronoi cells and at the center of each cell, one sensor is deployed. The work focuses on maximization of coverage area and uses deterministic deployment. F. Alassery [8] has proposed a node deployment design based on virtual multiple-input multiple-output technology to increase the performance of cluster based wireless sensor network. The height of the antenna of nodes is taken into consideration as an additional parameter for node deployment. This is helpful in increasing transmission range of nodes and minimizing relay nodes. X. Song et al. [9] have proposed a secure node deployment scheme based on evidence theory approach and caters for 3D underwater wireless sensor networks. This scheme implements sonar probability perception and an enhanced data fusion model to improve network coverage. It requires fewer nodes for large coverage area without compromising the quality of detection ability. Also it focuses on lifetime of the network. The cost of deployment is not considered here. S.M. Koreim and M.A. Bayoumi [10] have proposed a coverage hole detection algorithm for detecting coverage hole in wireless sensor network resulted due to damage of sensor nodes in area of interest due to flood or fire spreading. The algorithm partitions the area of interest into equal sized cells and each cell is cut into different triangles and then identifies the triangles that are not covered by any sensor. The triangles are formed with the help of three neighbouring sensor nodes. Energy efficiency here is achieved by minimizing the number of participating sensor nodes for hole detection. S. S. Kashi [11] proposed an algorithm named Heterogeneous Distributed Precise Coverage Rate (HDPCR) that detects holes and calculates coverage area of heterogeneous wireless sensor network using localized

mechanism. Boundary detection mechanism is used to determine the boundary of the hole area. A. Katti and D. K. Lobiyal [12] have proposed deterministic 3D node deployment strategy for wireless sensor network that includes prism deployment, pyramid deployment, cube deployment, hexagonal prism deployment for finding coverage prediction. It also determines the minimum number of sensor nodes required for specific coverage prediction. They have also proposed a scheduling algorithm for enhancing network lifetime. The work is suitable for small sized network. K. Wei [13] proposed a novel node deployment algorithm based on multi-objective evolutionary algorithm that optimizes average energy consumption, average sensitivity area and network reliability. In order to achieve the objectives, they have improved MOEA/D method by incorporating uniform design to generate aggregation coefficient vector and quadratic approximation for local search. The algorithm is designed for homogeneous network and no attempt is made to balance the energy efficiency and quality of operation.

N. Rai and R. D. Daruwala [14] have proposed an empirical formulation for estimation of randomly deployed nodes for attaining desired coverage for any size network. The formula includes the parameters that affect the optimum number of nodes. The formula is based on regression analysis using least square polynomial curve fitting technique. Sensor device characteristics are taken into consideration to devise the formula. Mainly, the authors have focussed on node density for desired coverage. S. Indumathi and D. Venkatesan [15] have proposed a dynamic node deployment model using genetic algorithm with gap cluster technique that uses different types of sensors. Gap cluster technique is used to determine the coverage hole resulted after initial deployment of nodes. In order to improve the network coverage and to minimize the number of gap clusters, additional nodes are deployed in gap region. The authors have not focused on the issue of energy efficiency and cost of the deployment. J. W. Lee and W. Kim [16] have proposed randomly deployed node deployment scheme that uses swarm intelligence for improving network lifetime and network coverage for heterogeneous wireless sensor network. This paper uses binary valued swarm intelligence algorithm such as Particle Swarm Optimization, Ant Colony Optimization, and Artificial Bee Colony Optimization. The work considers two types of nodes such as ordinary nodes and powerful nodes and focuses on minimization of network cost by minimizing number of nodes without compromising guaranteed coverage. The work is silent about how to handle the coverage hole problem in the network. M. R. Serik and M. Kaddour [17] have proposed a node deployment scheme for camera based wireless sensor network which focuses on optimization of deployment cost by minimizing the number of camera nodes required to cover a set of target objects with a pre-defined level of quality, position of camera nodes and orientation of camera nodes. Binary particle swarm optimization algorithm is used to minimize the number of camera nodes. The work is silent about energy efficiency of the network. Y. Yoon and Y. H. Kim [18] have proposed a node deployment algorithm based on the genetic algorithm for maximization of network coverage. A mixture of different types of static sensors is used for the deployment. The work focuses on the determination of

number of sensor of each type while maximizing network coverage. Network coverage is determined using Monte-Carlo method. When the sample size is very large or very small this way of network coverage calculation gives incorrect result. Z. Kang et al. [19] have proposed a decentralized, coordinate free, node based coverage hole detection algorithm which uses boundary critical points to determine hole and uses concept of perpendicular bisector for hole patching. The algorithm is suitable for grid type network and randomly deployed network. The objective of the algorithm is to achieve full coverage. S. Babaie and S. S. Pirahesh [20] have proposed a method that detects holes and their sizes in area of interest using voronoi diagram. Then holes are filled with mobile sensors. The issue of deployment cost and energy efficiency are not addressed. J. Wang et al. [21], proposed a PSO based energy efficient coverage control technique for homogeneous wireless sensor network in which the network in which the node locations are adjusted with respect to the coverage rate and energy consumption of each grid.

### III. ASSUMPTIONS, DEFINITION AND PROBLEM STATEMENT

#### A. Modelling Assumptions

- The area of interest is a two-dimensional plane area over which sensor nodes are randomly deployed.
- Sensor network is a heterogeneous sensor network.
- Sink knows the location of all nodes.
- Sensing region of a sensor is a circle.
- Cost of deployment is only based on number of nodes used for the deployment.
- Energy consumption is minimized by minimizing number of communication with the sink at the time of deployment.
- As network lifetime depends on energy consumption, by minimizing energy consumption lifetime can be increased.
- There is no obstacle in the network and the environment is noise free.
- Target object can be detected if it is in the sensing range.

#### B. Definitions

Let  $ST = \{1, 2, \dots, n\}$  are the static nodes and  $MB = \{1, 2, \dots, m\}$  are the mobile nodes are to be deployed on the area of interest. Out of  $m$  mobile nodes, some nodes are powerful mobile nodes and some nodes are ordinary mobile nodes. Area is divided into  $m1 \times n1$  grids and each cell size is  $d \times d$  where  $d$  is the diameter of sensing disk of static sensor (see Fig. 1). Let  $S = \{s_1, s_2, \dots, s_k\}$  is the set of  $k$  subsets, where  $k$  is the number of cells present in the area of interest (AOI). Each subset  $k_i$  consists of end points and center point of a cell:  $s_i = \{(x_{i1}, y_{i1}), (x_{i2}, y_{i2}), (x_{i3}, y_{i3}), (x_{i4}, y_{i4}), (c_{ix}, c_{iy})\}$ , where,  $(x_{ij}, y_{ij})$ ,  $j=1 \dots 4$  are the end points of the cell and  $(c_{ix}, c_{iy})$  is the center point of cell. A cell is individually analyzed to determine the hole.

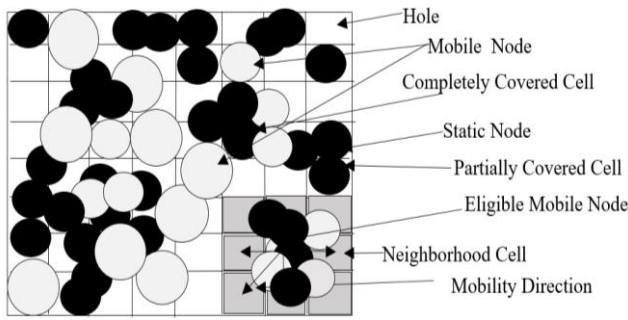


Fig. 1. Graphical Representation of Definition of Different Terminologies.

- Completely Covered Cell:-A cell  $i$  is said to be completely covered if all the points belonging to subset  $s_i$  is covered by sensors.
- Partially Covered Cell: - A cell  $i$  is said to be partially covered if at least one points belonging to subset  $s_i$  is not covered by any sensors and there is no intersection between the cell  $i$  and any of the sensor's sensing range.
- Uncovered Cell: - A cell  $i$  is said to be uncovered if none of the points belonging to subset  $s_i$  is covered by any sensors and there is no intersection between cell and sensing disk which is assumed as circle here.
- Hole:-It is the uncovered portion of cells.
- Hole Segment:-It is the area covered by adjacent holes.
- Eligible Mobile Nodes:-The mobile nodes whose sensing region is completely or partially overlapped with the sensing region of static nodes are said to be eligible mobile nodes.
- Neighborhood Cell:-Each cell is surrounded by eight equal sized neighborhood cell. A neighborhood cell of a cell may be uncovered, covered, partially covered by sensors.
- Optimum Location of Mobile Node:-It is the location that maximizes network coverage with minimum power usage if a mobile node is relocated to this location.
- Mobility Direction for Mobile Nodes:-An eligible mobile node can travel in any direction to cover the hole.
- Coverage Ratio:-It is the ratio between total covered cell and total number of cells present and its value lies in the range of  $[0, 1]$ .
- Detection Accuracy: - It is the ratio between the area under sensing coverage and the size of the AOI.

### C. Problem Statement

In a WSN designed for object detection and tracking, the full coverage of area of interest is required for continuous tracking of object. But when the nodes are deployed randomly across the region, coverage holes are created which leads to frequent failing of detecting the target even if the target is present. The matter worsens when the hole is present at the boundary of the area of interest and the size of the hole is

large. One of the solutions to it is to deploy additional nodes. Though, deployment of additional nodes helps to achieve full coverage but this limits network lifetime and increases network cost. Thus, there is a need of economic deployment plan that balances network coverage and network lifetime.

## IV. PROPOSED MODEL

### A. Optimum Number of Node Determination

This step deals with calculation of optimum number of nodes required to achieve desired coverage. Nodes deployed in the area of interest are a mixture of static and mobile nodes.  $M$  number of powerful mobile nodes and  $N$  number of ordinary nodes are used. One-third of  $N$  ordinary nodes are mobile nodes and remaining two-third nodes are static nodes. Ordinary nodes are with same sensing range, transmission range and battery power. Powerful mobile nodes are with high sensing range, transmission range and battery power in comparison to static nodes. The optimum number of powerful mobile nodes and ordinary nodes are determined using Particle Swarm Optimization (PSO) Based Technique [21, 22]. The PSO is a meta-heuristic optimization algorithm which is based on the behavior of the birds. The steps of the algorithm are given in Fig. 2.

```

1. Initialize Population and define search parameters
2. Initialize the velocity and position of the swarm. Also initialize best
   value of individual swarm ( $Pbest_i$ ) and global best ( $Gbest$ )
3. Find the fitness of the swarm using objective function
4. Update  $Pbest_i$  and  $Gbest$ 
5. Update the position and velocity of the swarm
The equation for updation of position and velocity are:

$$v_i(t+1) = w*v_i(t) + c_1*r_1*[Pbest_i(t) - x_i(t)] + c_2*r_2[Gbest(t) - x_i(t)]$$


$$x_i(t+1) = x_i(t) + v_i(t+1)$$

where,  $v$  is the particle velocity,  $x$  is the particle position,  $r_1$  &  $r_2$  are random numbers in the range of 0 and 1,  $c_1$  &  $c_2$  are learning factors.  $w$  is the inertia weight.
6. Repeat steps 3 to 5 until termination criteria is reached.
    
```

Fig. 2. Pseudo-Code of PSO Algorithm.

The objective function for node optimization is:

$$\text{Minimize } f(N, M) = N + M \quad \text{subject to}$$

$$C_{coverage}(N, M) \geq C_{guaranteed\_coverage}$$

$$\text{and } N > 0 \text{ and } M > 0 \quad (1)$$

Where,  $C_{coverage}(N, M)$  is the total area covered by  $N$  and  $M$  number of sensors.  $C_{guaranteed\_coverage}$  is the desired coverage area.

$$C_{guaranteed\_coverage} = \sum_{i=1}^N S_R + \sum_{i=1}^M S_{R1} \quad (2)$$

Where,  $S_R$  and  $S_{R1}$  are sensing range of ordinary sensor and powerful sensor respectively.

$$C_{coverage} = (1 - e^{-\pi\lambda})(1 - e^{-\pi\lambda}) \quad (3)$$

$$\text{Where, } \lambda = (N \times S_R^2) / \|A\| \quad (4)$$

Where  $\|A\|$  is the area of the monitoring area.



$$\lambda' = (M \times S_{R_1}^2) / |A| \quad (5)$$

$\lambda$  and  $\lambda'$  represents the probability of a point being covered by normal node and heterogeneous node respectfully.

$(1 - e^{-\pi\lambda})$  is the mean of the probability of covering a point by  $N$  normal nodes in the given area and  $(1 - e^{-\pi\lambda'})$  is the mean of the probability of covering a point by  $M$  heterogeneous nodes in the given area.

$$F(N, M) = \begin{cases} f(N, M), & \text{if } g(N, M) \geq 0, N > 0, M > 0 \\ f_{max} + |g(N, M)|, & \text{otherwise} \end{cases} \quad (6)$$

Where  $f_{max}$  is the worst fitness value of particles.  $g(N, M)$  is a normalized constraint which is calculated as follows

$$g(N, M) = \left( \frac{c_{coverage(N, M)}}{c_{guaranteedcoverage}} \right) - 1 \quad (7)$$

### B. Hole Determination

This step deals with the identification of hole in cells and calculating the size of the hole. The process of hole determination in a cell is treated here as the case of intersection of circle and rectangle. Each circle is a sensing disk and each rectangle is a cell in the area of interest. Let  $A, B, C, D$  are the end points of the rectangle.  $E$  and  $R$  are center and radius of the circle, respectively. A circle intersects a rectangle if the distance between the point of the rectangle closest to the center of radius is less than the radius of the circle. If  $(p_x, p_y)$  is an end point of the rectangle and  $(q_x, q_y)$  is the center of the circle and then, the Euclidean distance between them is:

$$distance = \sqrt{(p_x - q_x)^2 + (p_y - q_y)^2} \quad (8)$$

The hole area in the cell is determined using the formula of circle, sector, triangle and intersection of lines. If no end points are inside the circle and there is no intersection between circle and rectangle, then there is no overlapping between circle and rectangle and the rectangle is completely a hole. If all the end points are inside the circle, then circle has completely overlapped the rectangle. Different cases of intersection between and circle are:

- 1) All the end points of the rectangle are inside the circle.
- 2) Circle completely inside the rectangle.
- 3) Circle intersecting one side of rectangle.
- 4) Circle intersecting two adjacent sides of rectangle (including and excluding corner).
- 5) Circle intersecting two opposite side of the rectangle.
- 6) Circle intersecting three sides of the rectangle.
- 7) Circle intersecting four sides of the rectangle.

Following is an example of calculation of hole in a rectangle when circle intersects one side of rectangle. See Fig. 3(a). Let  $L$  and length and  $B$  is breadth of rectangle,  $\theta$  is the angle of sector and  $R$  radius of the circle, and radius of both circles are same then,

$$\text{Hole in ABCD rectangle} = (L \times B) - \text{Area of shaded region} \quad (9)$$

For Fig. 2(a),

$$\text{Area of shaded region} = 0.5 \times R^2 (\theta - \sin \theta) \quad (10)$$

Where,  $\theta$  is the angle between two sides of the sector.

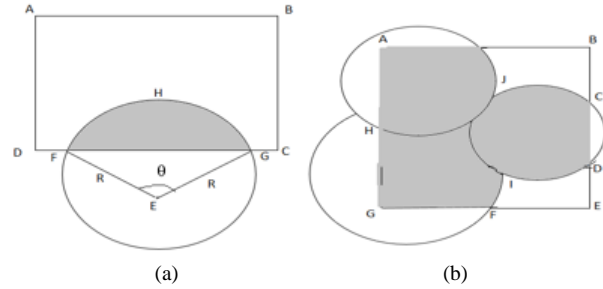


Fig. 3. (a)-(b): Intersection of Circle and Rectangle.

For the case of Fig. 3(b), individual case of overlapping is determined and then circle-circle intersection area is determined to calculate the hole region in the rectangle. The area of overlapped portion is the difference between area of rectangle and (sum of the shaded portion of circle-overlapped area of circle). The area of overlapping  $AO$  between two circles for circles having two different radius  $r$  and  $R$  is calculated as follows:

$$AO = r^2 \times \text{atan2}(t, d^2 + r^2 - R^2) + R^2 \times \text{atan}(t, d^2 - r^2 + R^2) - \left(\frac{t}{2}\right) \quad (11)$$

Where,  $d$  is the distance between center of two circles and

$$t = \sqrt{(d + r + R)(d + r - R)(d - r + R)(-d + r + R)} \quad (12)$$

If  $t$  does not contain any imaginary part, there is no intersection between circles.

There may be single hole region or multiple hole region in a rectangle. The area of hole in a rectangle is determined as:

$$Hole_{Area_i} = \sum_{j=1}^k A_j \quad (13)$$

Where  $k$  is the no of hole regions present in the  $i_{th}$  rectangle. A hole segment consists of adjacent hole regions (see Fig. 4).

$$Hole_{AOI} = \sum_{j=1}^{m1 \times n1} A_j \quad (14)$$

Where  $Hole_{AOI}$  is the total hole area in area of interest and  $m1 \times n1$  are the number of cells in area of interest.

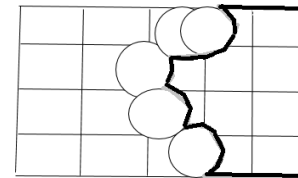


Fig. 4. Hole Segment (Black Marked Region).

### C. Hole Filling

Ordinary mobile node movement is restricted to the hole region of neighborhood cell whereas powerful mobile nodes can travel optimum distance to cover the hole area. By doing so, energy consumption can be minimized at the time of node redeployment and network coverage can be maximized. First, the list of eligible nodes for relocation is prepared. In the first phase, eligible ordinary mobile node movement is done and

they are moved to the optimum location of the neighborhood. The node must move at most  $d1$  distance towards neighboring hole region to maximize coverage. See the Fig. 5(a),  $A$  and  $B$  are the center of two circles respectively.  $C$  and  $D$  are the intersection point of two circles.  $AG$  and  $BF$  are the radius of first and second circle respectively. The value of  $d1$  is calculated as follows:-

$$d1 = FG + rand \quad (15)$$

Where  $FG = FE + EG$  and  $rand$  is a random number in the range of  $[0, 1]$ .  $FE$  and  $EG$  can be calculated using the equation for calculating the area of the circular segment in a sector of a circle. New location in the neighborhood is shown in Fig. 5(b). Hole area in each concerned cell is updated after relocation of ordinary mobile nodes.

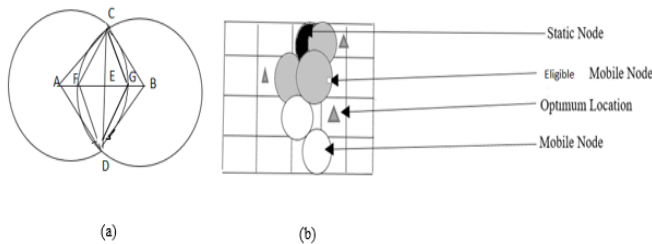


Fig. 5. (a) Width of Overlapping Sensing Area between Two Nodes (b) Optimum Location for Relocation of Ordinary Mobile Node.

In second phase, the optimum location of the eligible powerful nodes is decided for relocation in order to increase the network coverage. PSO algorithm is used to determine the optimum location of these mobile nodes. The objective function here used is

$$f(x_{new}, y_{new}) = ((x_{old} + U_{u1}), (y_{old} + U_{u2})) \quad (16)$$

$$\text{Subject to } ((\sum_{i=1}^N \pi r^2 + \sum_{j=1}^M \pi R^2) - A_{ov}) > TC$$

$$0 > U_{u1} \leq X \text{ and } 0 > U_{u2} \leq Y$$

Where  $(x_{new}, y_{new})$  is the new location of eligible powerful mobile node,  $(x_{old}, y_{old})$  is the current location of it.  $U_{u1}$  and  $U_{u2}$  are the amount of change required in x-direction and y-direction, respectively.  $TC$  is the area coverage currently and  $A_{ov}$  is the overlapping area.  $X$  and  $Y$  define AOI size.

$$U_{u1} = \min(dif f_{x1}, dif f_{x2}, dif f_{x3}, \dots, dif f_{xk}) \quad (17)$$

$$U_{u2} = \min(dif f_{y1}, dif f_{y2}, dif f_{y3}, \dots, dif f_{yk}) \quad (18)$$

Where  $dif f_{xi}$  is the distance between x-coordinate of node and the x-coordinate of centroid of hole  $I$ ,  $dif f_{yi}$  is the distance between y-coordinate of node and the y-coordinate of centroid of hole  $i$ .

The flowchart of the proposed model is given Fig. 6 and its corresponding algorithm is described below:

1. Initialize the AOI size
2. Determine number of static nodes, ordinary mobile nodes and powerful mobile nodes required for desired coverage in AOI.
3. Initialize sensing radius of all sensors

4. Deploy all nodes randomly
5. Divide AOI into grids and define coordinate of each cell.
6. For  $i=1$  to number of cells //Hole Detection and its size //calculation  
 $NS=0$ ;  
 $Cell_i = \text{End points of cell } i$   
 For  $k=1$  to  $\text{size}(Cell_i)$   
 For  $j=1$  to number of sensors  
 Calculate  $d(j)$  for all end points of  $i$ th cell  
 If  $d(j) < \text{sensing range}$   
 Update  $NS$  by 1.  
 End  
 End  
 End  
 For  $j=1$  to number of sensors  
 Find the intersection between  $i$ th cell and  $j$ th sensor  
 End  
 If  $NS=0$  and there is no intersection between sensor and cell  
 $Cell\_status(i) = \text{'Hole'}$   
 Store the Hole area size  
 Else if  $NS=4$   
 $Cell\_status(i) = \text{'Completely Covered'}$   
 Else if  $NS < 4$  and there is overlapping between sensor and cell  
 $Cell\_status(i) = \text{'Partially Covered'}$   
 End  
 If  $Cell\_status(i) = \text{'Partially Covered'}$   
 Identify hole regions and calculate their size  
 Find the sum of all hole regions  
 End  
 End  
 7. Calculate the total Hole size in AOI  
 8. For  $j=1$  to number of mobile nodes //Hole Filling  
 Determine the eligible mobile nodes for relocation  
 End  
 9. For  $i=1$  to number of eligible ordinary mobile nodes  
 Determine the new location for nodes in neighborhood  
 End  
 10. For  $i=1$  to number of eligible powerful mobile nodes  
 Determine the new location for nodes in AOI  
 End  
 11. Calculate Hole size in AOI

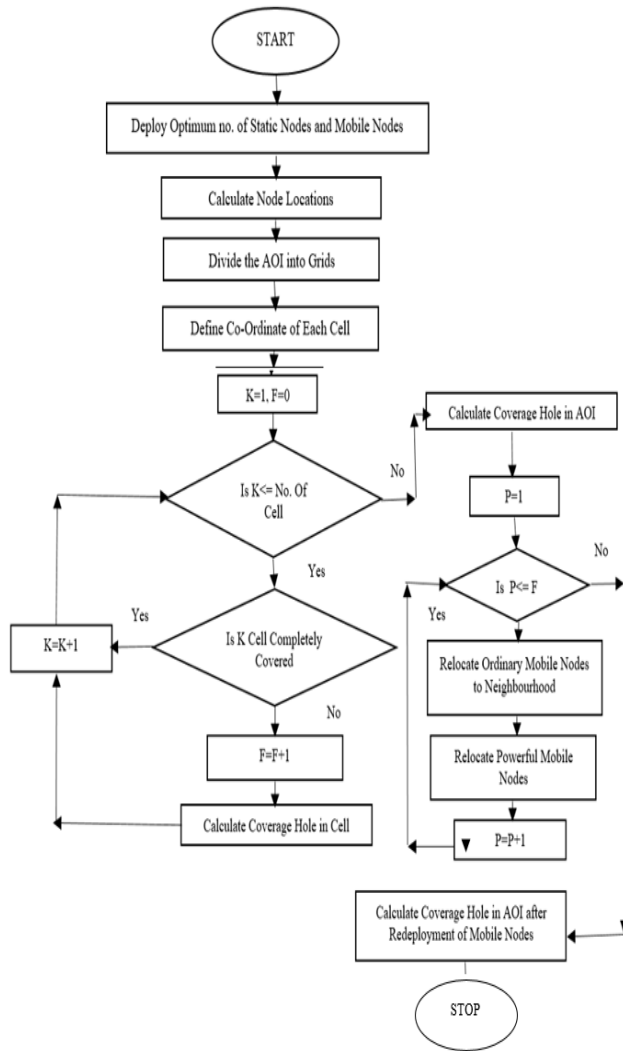


Fig. 6. Flow Chart of Proposed Model.

#### D. Energy Model

For static node, energy consumption is negligible at the time of deployment and hence, the consumption value is neglected. In case of mobile nodes, it is based on the distance travelled by the node and the communication done with sink. If each packet size sent to the sink is  $k$ -bit and each node sends one packet to sink and receives one packet from the sink, then the power consumption for sending and receiving packet from sink, the following set of equation can be used to calculate the energy consumption by the mobile node.

$$E_{TX}(k, d) = \begin{cases} k * E_{elec} + k * E_{amp} * d^2, & \text{if } d < d_0 \\ k * E_{elec} + K * E_{fs} * d^4, & \text{if } d \geq d_0 \end{cases} \quad (19)$$

Where  $E_{TX}$  the energy required for packet transmission,  $d$  is the distance between source and destination and  $d_0$  is the threshold distance.  $E_{elec}$  is the base energy which required to run the transmitter or receiver.  $E_{fs}$  and  $E_{amp}$  are the unit energy required for transmitter and amplifier.

$$\text{Where, } d_0 = \sqrt{\frac{E_{fs}}{E_{amp}}} \quad (20)$$

The energy required for receiving  $k$ -bit message is:

$$E_{RX}(k) = k * E_{elec} \quad (21)$$

Let  $E_{RELOC}$  is the energy required to relocate a node from one location to other location situated at distance  $d$ .  $E_{RELOC}$  depends on the distance between old location and new location of the same node. Then the total energy consumed by the node for movement from one location to other is:

$$E_c = E_{TX}(k, d) + E_{RX}(k) + E_{RELOC} \quad (22)$$

## V. PERFORMANCE EVALUATION

### A. Simulation Environment and Parameter Setup

The proposed model is simulated using MatLab R2020b. The AOI size is 500m×500 m. The network is a heterogeneous wireless sensor network and nodes are deployed randomly. All static and ordinary mobile nodes are with same sensing radius and same battery power. The only difference is the mobility. The sensing radius and battery power for this category homogeneous node are 20m and 5J respectively. To reduce the power consumption in case of ordinary mobile nodes during node deployment, their mobility is restricted to short distance. Powerful mobile nodes are with sensing radius of 30 m and battery power of 10J. The communication range in case of all nodes are the twice of their sensing radius. The location of the sink is [500,500]. The total number of nodes required for full coverage of AOI is calculated using PSO algorithm i.e. the number of same characteristic nodes  $N$  and powerful nodes  $M$  are optimized. One-third of  $N$  nodes have the mobility. The coverage degree  $k=1$ . Total 60 static nodes, 30 ordinary mobile nodes and 46 powerful mobile nodes are used for deployment. For the movement of ordinary mobile nodes, a 3×3 mask is used and 8-neighborhood cells of the cell containing overlapped mobile node is determined. The ordinary mobile node is moved to the hole present in the neighborhood. For the powerful nodes, PSO algorithm is used to determine the optimum location. Algorithm runs for 200 iterations. As one of the factor that affects the network cost is the number of nodes used for deployment, we assume that the cost can be minimized if the number of hardware to be used is minimized. Also, when the number of nodes is minimized, number of communication between sink and nodes are minimized. Hence, Energy is saved. Following are the values of PSO parameter:  $c1=c2=2$ ,  $0.4 \leq w \leq 0.9$ ,  $v=0.1 \times \text{InitialPosition}$ , number of swarm=100. The values of parameters for harmony search algorithm are:  $\text{hms}=5$ ,  $\text{hmcr}=0.95$ ,  $\text{par}=0.25$ ,  $\text{bw}=0.2$ ,  $\text{numRows}=30$ .

### B. Simulation Result

Fig. 7(a) shows the random deployment of nodes in AOI resulting multiple coverage hole and Fig. 7(b) scenario after hole filling. The quality of operation in a network mainly depends on network coverage and lifetime of the network. There is a direct relationship between lifetime of the network and network coverage. Also, when an object is in the hole region is not detected by any sensor. Object's movement is simulated and the moving path is a simple straight line. At every 't' time interval, a sensor searches for presence of an object in its sensing coverage area.. So, the detection accuracy depends on the maximum area coverage. Hence, we define the

object detection accuracy in terms of sensing area coverage. Fig. 7(b) shows the status of AOI after moving mobile nodes to the hole region. Fig. 8(a) and 8(b) conveys that the network coverage increases when the overlapping area between sensors are decreased. Fig. 8(c) shows the energy consumption during movement of mobile node. Table I shows the performance of the proposed algorithm for varying area size with constant sensing range of nodes.

In Table II, area size(AS) is  $500 \times 500 \text{ m}^2$ .  $SR_1$  is the sensing range of static sensor and ordinary mobile sensor. NS is the number of static node, NOM is the number of ordinary mobile node and NOP is the number of powerful mobile nodes. CBD is the coverage before deployment and CAD is the coverage after deployment.  $SR_2$  is the sensing range of powerful mobile sensor. Coverage ratio(CR) is the ratio of number of cell fully covered by total number of cells. From the Tables I and II, it is clear that coverage ratio depends on the on the location of sensor nodes in AOI. Fig. 9(a) and 9(b) shows the performance comparison between the proposed model and other existing model. The number of nodes shown in the Fig. 9 is for different area sizes mentioned in Table I. From Fig. 9(a) and 9(b), it is clear that coverage ratio and energy efficiency of proposed model is better in comparison to HSA and PSO algorithm. Both HSA and PSO based node deployment technique are silent about the effect of node density on energy efficiency and cost of the deployment.

Using static nodes only for deployment makes the system cheaper but creates a lot of coverage hole in AOI which greatly affects the quality of operation, particularly when the network is used for object tracking application. Coverage holes resulted the missing of object and hence increases the rate of reporting the false negative message about object's presence. It is obvious that when coverage hole minimizes, object detection rate increases under the assumption that the environment is noise free and the signal attenuation is in the tolerable limit. The detection accuracy is directly proportional to the coverage area. Proposed model achieves 97% accuracy in the case of no obstacle in the monitoring area and the surface area is a 2d-plane.

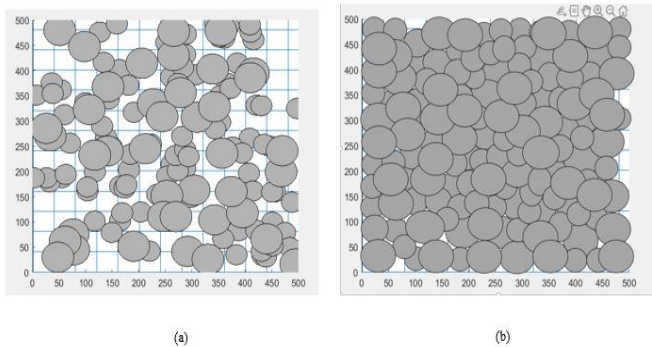


Fig. 7. (a) Node Deployment before Node Location Optimization (b) Node Deployment after Node Location Optimization.

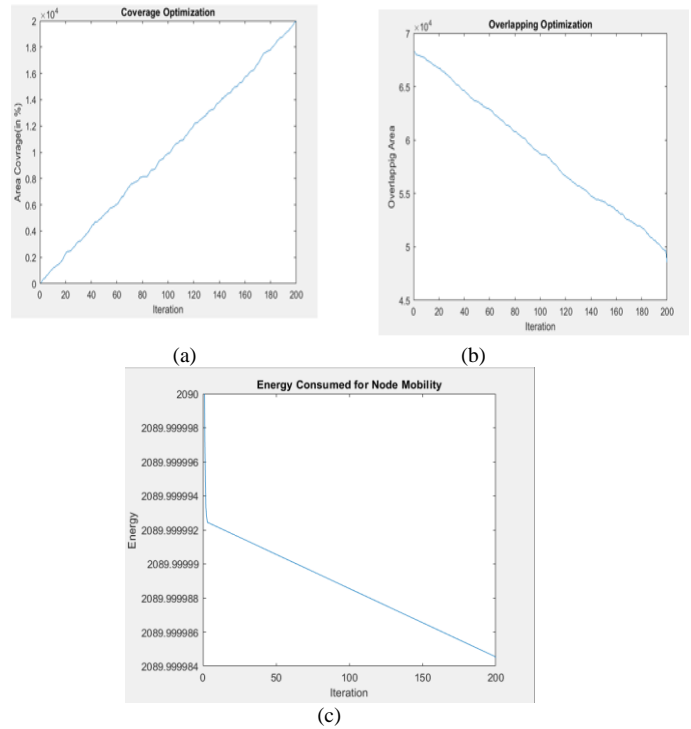


Fig. 8. (a) Sensing Area Coverage Optimization (b) Overlapping Area Optimization (c) Energy Consumed for Node Movement of Proposed Model.

TABLE I. PERFORMANCE OF PROPOSED MODEL FOR DIFFERENT AREA SIZE FOR CONSTANT SENSING RANGE

AS (in $\text{m}^2$ )	NS	NOM	NOP	CBD (in %)	CAD (in %)	CR
$500 \times 500$	60	30	40	76.6760	96.9956	0.9100
$400 \times 400$	59	30	15	61.7031	91.9300	0.9400
$300 \times 300$	59	30	15	80.9911	93.3933	0.9531
$350 \times 350$	53	27	14	78.5889	93.3933	0.9531
$200 \times 200$	47	24	14	80.2300	91.5622	0.9531

TABLE II. PERFORMANCE OF PROPOSED MODEL FOR SAME AREA SIZE WITH DIFFERENT SENSING RANGE

$SR_1$ (in m)	$SR_2$ (in m)	NS	NOM	NOP	CBD (in %)	CAD (in %)	CR
20	30	60	30	40	74.6760	96.9956	0.9100
15	30	68	34	54	51.1996	94.1908	0.8713
25	30	56	29	25	40.6496	92.2032	0.8260
20	25	63	31	55	57.2932	94.9696	0.8505
15	20	115	57	82	75.8592	96.4364	0.9273

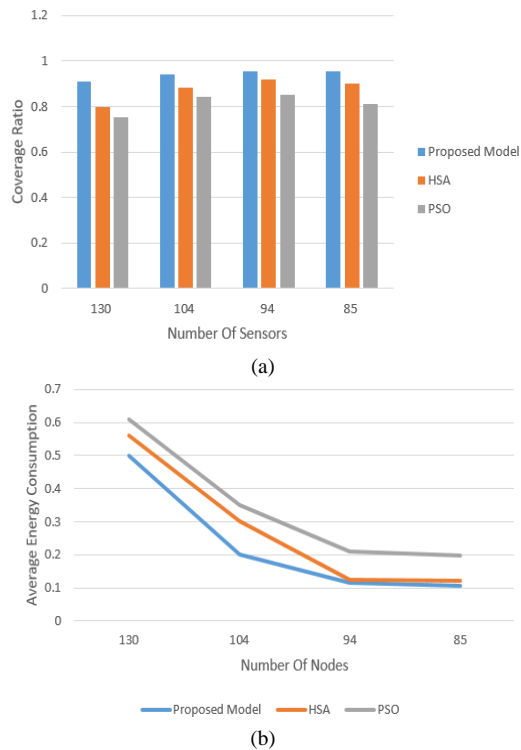


Fig. 9. Performance Comparison of Proposed Model with Other Model (a) Coverage Ratio (b) Average Energy Consumption.

Using mobile nodes only for deployment achieves high quality of operation by achieving full coverage but makes the system costlier. So, the proposed model tries to balance between cost of the system with quality of operation by considering both type of nodes for deployment. To calculate the deployment cost, we have considered the hardware cost only. Let for deploying a static node, deployment cost is 1 unit of money, for ordinary mobile node it is 1.5 unit of money and 2 unit of money for powerful mobile node, then, the deployment cost for an area of size  $500 \times 500 \text{ m}^2$  is 185 unit of money.

## VI. CONCLUSION

Network coverage and connectivity is the one of the factor on which the object detection accuracy in WSN depends. Coverage hole problem is a common problem in a randomly deployed network. Presence of object in hole region cannot be detected and necessary action cannot be initiated. The full coverage of AOI is desired which can only be achieved by deploying additional mobile nodes in hole region. It is exactly difficult to calculate that how many mobile nodes will be deployed. A large number of mobile nodes increases the deployment cost and shortens network lifetime whereas the small number of mobile nodes will not be able to fill the hole. In our work, we have tried to balance the number of mobile nodes with network cost and energy. We assume that when the number of nodes is minimized, number of communication can be minimized which in turn will increase energy efficiency. Proposed model uses a mixture of static nodes and two types of mobile nodes. Number of static nodes and mobile nodes for deployment are optimized using PSO algorithm. After initial deployment, the overlapped ordinary mobile nodes are shifted

to the hole region present in the neighborhood. For determination of destination for them, a  $3 \times 3$  mask is used and 8-neighborhood cells are determined. Nearest destination is the new location of the ordinary mobile node. For powerful mobile nodes, the new location is determined using PSO algorithm. The performance of the proposed model is evaluated and its coverage ratio is 0.91 and coverage percentage is 96.9956. Its performance is compared with other existing algorithms and the proposed model is found better in comparison to other. The detection accuracy for the proposed model is 97%. In future, we will try to focus on other artificial intelligence techniques that can further improve the detection accuracy by improving the coverage ratio. In this work, we have considered the disk model of sensing for coverage calculation. We will also focus in other model of sensing for the same.

## REFERENCES

- [1] M. Farsi, M. A. Elosseini, M. Badawy, H. A. Ali and H. Z. Eldin, "Deployment Techniques in Wireless Sensor Networks, Coverage and Connectivity: A Survey"; IEEE Access, vol. 7, February 2019, pp. 28940-54.
- [2] S. Commuri and M. K. Watfa, "Coverage Strategies in Wireless Sensor Networks", International Journal of Distributed Sensor Networks, vol. 2, pp. 333-353, 2006.
- [3] B. A. Fuhaidi, A. M. Mohsen, A. Ghazi and W. M. Yousef, "An Efficient Deployment Model for Maximizing Coverage of Heterogeneous Wireless Sensor Network Based on Harmony Search Algorithm", Hindawi, Journal of Sensors, vol. 2020, pp. 1-18.
- [4] C. Zygowski and A. Jaekel, "Optimal Path Planning Strategies for Monitoring Coverage Holes in Wireless Sensor Networks", Ad Hoc Networks, vol 96, January 2020, pp. 101990.
- [5] I. Alablani and M. Alenazi, "EDTD-SC: An IoT Sensor Deployment Strategy for Smart Cities", Sensors, December 2020, vol. 20, issue 24, pp. 7191-8010.
- [6] P. prabhakaran, R. Jayavade, L. Malathi and M. Ramesh, "Energy Efficient Object Tracking Using Adaptive Node Deployment and Evolutionary Algorithm Based Node Localization", International Journal Of Scientific & Technology Research, November 2019, vol. 8, issue 11, pp. 1960-69.
- [7] J. Mao, X. Jiang and X. Zhang, "Analysis of Node Deployment in Wireless Sensor Networks in Warehouse Environment Monitoring Systems", Eurasip Journal on Wireless Communications and Networking, December 2019, vol. 2019, issue 1, pp. 1-15.
- [8] F. Allassery, "Convergence between Virtual MIMO and Node Deployment Strategy for High Performance Multi-hop Wireless Sensor Networks", IAENG International Journal of Computer Science, June 2019, vol. 46, issue 2, pp. 349-357.
- [9] X. Song, Y. Gong, D. Jin and Q. Li, "Nodes Deployment Optimization Algorithm based on Improved Evidence Theory of Underwater Wireless Sensor Networks", Photonic Network Communications, 2019, vol. 37, issue 2, pp. 224-232.
- [10] S.M. Koreim and M.A. Bayoumi, "Detecting and Measuring Holes in Wireless Sensor Network", Journal of King Saud University-Computer and Information Sciences, 2018, vol. 32, no. 8, pp. 909-916.
- [11] S. S. Kashi, "Area Coverage of Heterogeneous Wireless Sensor Networks in Support of Internet of Things Demands", Computing, 2018, vol. 101, no 4, pp. 363-385.
- [12] A. Katti and D.K. Lobiyal, "Node Deployment Strategies and Coverage Prediction in 3D Wireless Sensor Network with Scheduling", Advances in Computational Sciences and Technology, 2017, vol. 10, no 8, pp.2243-2255.
- [13] K. Wei, "Node Deployment for Wireless Sensor Networks Based on Improved Multi-objective Evolutionary Algorithm", International Journal of Internet Protocol Technology, 2017, vol 10, no. 3, pp 189-195.
- [14] N. Rai and R.D. Daruwala, "Empirical Formulation for Estimation of Optimum Number of Randomly Deployed Nodes in WSN", International

- Conference on Wireless Communications, Signal Processing and Networking (WISPNET), 2016, pp 376-380.
- [15] S. Indumathi and D. Venkatesan, "Improving Coverage Deployment for Dynamic Nodes using Genetic Algorithm in Wireless Sensor Networks", Indian Journal of Science and Technology, 2015, vol. 8, no. 16, pp. 1-6.
- [16] J.W. Lee and W. Kim, "Design of Randomly Deployed Heterogeneous Wireless Sensor Networks by Algorithms Based on Swarm Intelligence", International Journal of Distributed Sensor Networks, 2015, vol. 11, no 8, pp. 1-8.
- [17] M. R. Serik and M. Kaddour, "Optimizing Deployment Cost in Camera-Based Wireless Sensor Network", IFIP International Conference on Computer Science and its Application, 2015, pp. 454-464.
- [18] Y. Yoon and Y. H. Kim, "An Efficient Genetic Algorithm for Maximum Coverage Deployment in Wireless Sensor Network", IEEE Transactions on Cybernetics, 2013, vol. 43, no 5, pp. 1473-1483.
- [19] Z. Kang, H. Yu and Q. Xiong, "Detection and Recovery of Coverage Holes in Wireless Sensor Networks", Journal of Networks, April 2013, vol.8, no.4, pp. 822-828.
- [20] S. Babaie and S. S. Pirahesh, "Hole Detection for Increasing Coverage in Wireless Sensor Network Using Triangular Structure", International Journal of Computer Science Issues, January 2012, vol. 9, issue 1, no.2, pp. 213-218.
- [21] J. Wang , C. Ju , Y. Gao , A. K. Sangaiah and G. J Kim, "A PSO based Energy Efficient Coverage Control Algorithm for Wireless Sensor Networks", Comput. Matter. Contin. January 2018, vol. 56, no.3, pp. 433-446.
- [22] A. P. Laturkar and P.Malathi, " Coverage Optimization Techniques in WSN using PSO: A survey", International Journal of Computer Applications, 2015, vol. 975, pp. 19-22.



# Fine-grained Access Control in Distributed Cloud Environment using Trust Valuation Model

Aparna Manikonda<sup>1</sup>

Research Scholar, Department of Computer Science  
Nitte Meenakshi Institute of Technology, Karnataka, India

Nalini N<sup>2</sup>

Professor, Department of Computer Science  
Nitte Meenakshi Institute of Technology, Karnataka, India

**Abstract**—Cloud computing has been in existence as an adaptable technology that gets integrated with IoT, Big-Data, and WSN to provide reliable, scalable and mesh-free services. However, because of its openness in nature, the privacy of the cloud is an important parameter for today's research. The most important privacy factor in cloud is access control and user trust. Many articles related to access control and trust management were presented, but most of them include highly complex algorithms that result in network overhead. This proposed security framework is for a better and more effective system wherein multiple distributed centers are created with trust-based computing for authentication and validation of requests from users and their resources. The idea of trust here is for efficient decision-making and establishing reliable relationships among users and resources using least computations. Each user has different permissions for each file present in the cloud server. The simulated results shows improvement in the rate of successful transactions, time cost and network overhead.

**Keywords**—Fine-grained; distributed; access control; trust

## I. INTRODUCTION

The security issues like privacy, trust, authentication and authorization need more attention with the rapid advancement in day-to-day technologies. Among them, access control and trust management are critical and complex issues that require more focus. But in the cloud environment, the access control approaches are semi-trusted because of the users snooping nature [1-2] that offer the resource or its attributes a complete access based on the rights of the user. Most of the access management methods [3-7] use encryption and decryption algorithms for protection of legit users. To reduce the computation overheads, many researchers [8-12] used the trust parameter in the process of decision-making for validation and authentication of user and their resources.

The research issues subjected to existing techniques involve the following:

- 1) The methods related to fine-grained involve lot of mathematical computations.
- 2) Trust-based involves subjective assignment of weights to the attributes considered for calculation of trust value which leads to time cost, and
- 3) Centralized-based leads to network overhead.

Hence, the designed model is named as distributed fine-grained access control using trust management (DFGACT). In this work, multiple distribution centers are created for

accessing the data by authorized users on basis of their trust values associated with them. Rest of the paper is organized as follows: Section II is literature survey, Section III is proposed approach and Section IV is results and analysis about the proposed approach followed by a conclusion in Section V.

## II. RELATED WORK

One of the mostly used cryptographic access control method is designed by Sahai [13] for volatile cloud situations. Many literature papers used CP-ABE and KP-ABE schemes to secure data processing in the cloud and WSN [14-18]. In these methods, the complexity of encryption growth is linear with the count of each attribute and conveys heavy computation overheads. CP-ABE schemes are the most used for fine-grained security in cloud computing. The attribute statistics [19] are completely hidden inside the access policy by way of the use of the randomizable fuzzy approach for decryption purposes. Somchart et al. [20] used the CP-ABE method for mobile cloud environments by way of introducing new proxy encryption to reduce the cost of decryption and encryption for mobile users. But the outsourcing encryption isn't always specified. CP-ABE calls for data proprietors to generate multiple ciphertexts which result in sizable overheads in computation. To triumph over this, an LSSS based CP-ABE has been proposed [21] that can decrypt the records that are relatable with this matched part.

Anil Kumar et al. [22] tried to triumph over the troubles associated with RBAC, where users can access entire object without any further authentication once access is granted by manipulating swift storage. Qian et al. [23] proposed a Merkle tree based on time and attribute that stores private keys of the user for decryption purposes to efficient access of the resources. A lightweight statistical computation [24] is carried out by the cloud server for granting unique access privileges to individual users. However, with the increasing variety of attributes the overhead increases.

The large statistics are stored specifically in the cloud for controlling the access of a massive amount of data with closed permissions of individual users [25]. However, this scheme hides total attributes of a user for getting an entry pass to the access rights. The conventional cloud storage structures goes through a hassle of returning the incorrect seek consequences or not going back to the total seek results, this can be solved by using the applied decentralized system model with the cipher textual content-based key-word seek characteristic according to the smart reduced in size Ethereum blockchain [26].

Qi Li et al. [27] constructed a scheme named SEMAAC for mHealth applications that have IoT enabled to achieve the functions of de-centralization. Here, each CSP computes the decryption costs by interacting with Associated Authority. But, the verification time of each PDC increases linearly with the number of associated authorities. For a specific biological system of healthcare industry in cloud computing [28] the user's closed permissions are mixed with fog computing to provide high-level security.

Roy et al. [29] constructed a concept in the direction of transportable cloud computing for the healthcare industry. This portable cloud display helps in analyzing the statistics with recognition of the patients' records and in removing proposals in medicinal services programs. A dynamic authorized system was proposed for cloud computing [30] where the authorized users are identified based on trust value. The access privileges of these users are created on their behaviour with the system. However, this work adds an overhead due to its various assignments and removal of permissions for malicious users.

### III. PROPOSED WORK

There are three important issues while accessing data from the server. Firstly, users having access to servers may additionally try and access data that isn't always intended for them. To keep away from such problems, every user needs to have a particular privilege to access the server. Secondly, Trust is an important parameter for improving the relationship between the user and their resources. Thirdly, a maximum of the theories mentioned for improving safety features are liable to a single point of failure. However, the decentralized version has higher throughput and lower fees together with overcoming the failure of single-factor issues within the system as compared to the traditional cloud storage device.

The main idea behind this research is to eliminate the complex computations that are involved in the existing state of art methods. The novelty behind this access-control method lies in the combination of 1) Decentralized approach to avoid single-point failures 2) Fine-grained approach for unique privileges for users and 3) Trust for an effective authorization process. Although many methods related to the above discussed issues are existing but none of the techniques is a combination of these three tactics. The reason to combine these three tactics is for better time cost, success rate and reduction of network overhead.

This phase of work defines a simplification of the proposed scheme DFGACT. Requirements in this approach are:

- 1) Multiple VMS of cloud that acts as a distribution center for the scheme.
- 2) Each DC has a cloud service provider that takes care of the incoming requests from the user and owner of the file/resource.
- 3) The distribution center stores the consumer/User information, trust matrices and permission table of the resource.
- 4) Each file has a permission table, the owners of the file/resource decide on the permissions that a user/consumer

should have and accordingly receives the PID of the file/resource that authorize them to access records.

Here, the user/consumer request is processed through these modules as shown in Fig. 1 to undergo the authorization and authentication system. The consumer request is of the form (UID, PID, RID, Trust value). The representation for the same is given in the table 1. To gain access control, the method used an idea similar to [31]. The access policy is a fine-grained in nature; this combines three methodologies RBAC, ARL and ABAC. For fine-grained precision, a set of access rules is assigned to each user that defines the access rights of that resource/file. In the proposed case, each file has a set of permissions in the form of attributes. The values associated with the attribute decide the right to access the resource/file. Every file/resource and a unique permission ID has been assigned to the user based on the initial demand/request. No two users can have the same permission ID for that resource.

However, the uniqueness of the work is addition of distribution centers that distributes trust matrices among neighboring CSPs for identification of valid users to avoid malicious requests for a better network overhead. Besides, trust computing is taken into consideration as a measurable factor in the scheme that considers users trust credentials and computing power of resources for better success rate. Table I describes the notations used in the paper.

TABLE I. NOTATIONS

Sl.no.	Notation	Meaning
1	AC	Authorization Centre
2	DC	Distribution Centre
3	DO	Data Owner
4	CNM	Data Consumer
5	ALC	Access level code or permission IDs
6	CSP	Cloud service provider
8	$C_a$	Type of user
9	$C_t$	Authentication degree
10	$T T_{VAL}$	Total Trust value
11	$T_{THRS}$	Trust threshold
12	$M$	Unique code in the permission table
13	$N$	Set of permissions associated with FILE
14	$Z$	File name
15	$RQ$	Degree of request potentiality
16	$r$	Priority of current request and is decided by CSP of DC
17	$RQ_f$	Number of times the user requested for the resource
18	$RQ_s$	Degree of valid request.
19	$R_{id}$	Request ID
20	$(U_{id})$	USER ID
21	$P_{id}$	Permission ID
22	$(F_{id})$	FILE ID

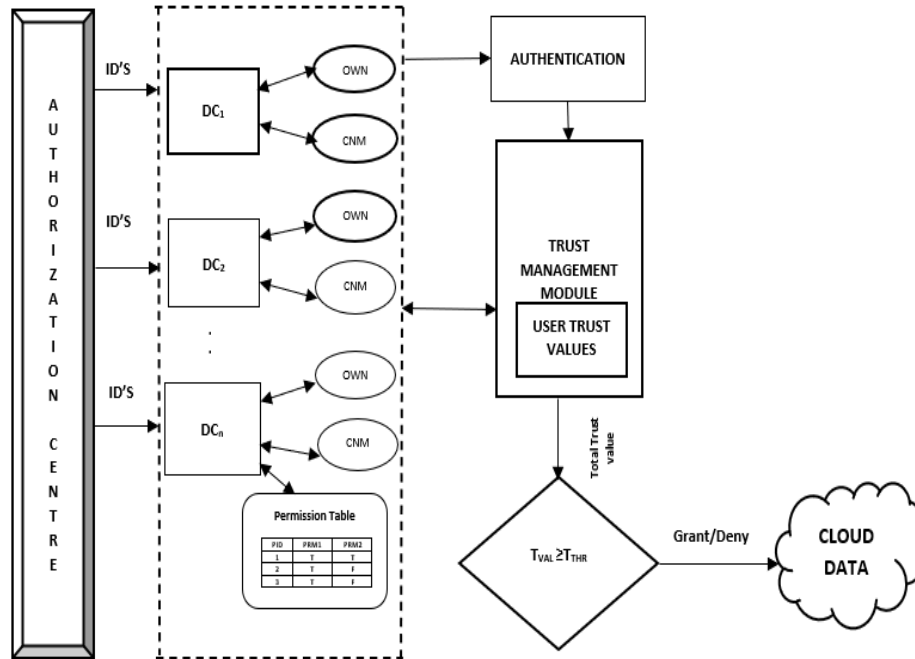


Fig. 1. System Architecture of DFGACT.

The scheme starts with an authentication procedure which involves the verification of user identity by its necessary associated credentials, such as User-Id, password, and RSA token. For the valid user, the activity is been recorded by the system in a log file for future reference. The request gets dropped for an invalid user and the information about the same is sent to remaining distribution centers. Followed by an authorization procedure, where the DC extracts the permission ID of the resource from the user request and assign the data to client according to the permissions associated with the  $F_{id}$  ( $P_{id}$ ).

#### A. System Architecture

The proposed architecture distributed fine-grained access control with trust computing named DFGACT includes a trust management module and an access control module in conjunction with four entities as shown in Fig. 1. The entities are Authorization center (AC), Distribution Center (DC), Data Owner (DO), and Data Consumer (CNM). In the proposed scheme, the consumer can get the right of entry to the resource through the ALC code of the report/ resource. Each of the entities with its functionalities is mentioned as underneath:

**USER/Owner:** This entity can store or access the information; the one tries to read /write the resource and this must be a valid user before accessing the resource.

**CSP:** This entity provides services to the authorized users. This resides inside the distribution centre.

**Distribution Centre (DC):** Each distribution center generates ID'S to user and owner of the resources. This entity stores the permission table of each resource, access policy, information about the owner and consumer. Every distribution

center has a CSP provider to validate the authorized users for accessing the resources.

**Authorization Centre (AC):** Authorization center creates various distribution centers. A unique ID is associated for each distribution center. This entity acts as a database for storing information about the distribution center.

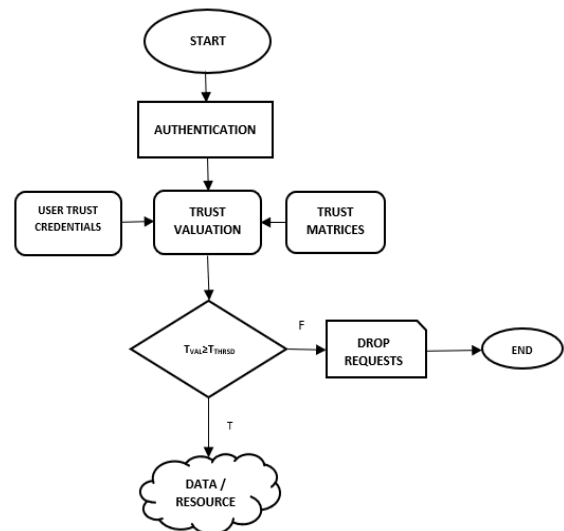


Fig. 2. Flowchart of the Proposed Architecture DFGACT.

The working principle of DFGACT is as under and the flowchart for the same is shown in Fig. 2:

- 1) Request from USER or Consumer to get entry into the Cloud data or Services through web interface of a specific distribution centre.

2) The distribution centre authenticates every individual via analysing the important related credentials such as  $C_a$  to identify the user is a trusted or untrusted. If the  $C_a$  are determined legitimate, then the request for access undergo trust evaluation process, or else the request of the consumer is denied for any additional processing.

3) If authentication check fails the CSP in that particular DC's drops the request and transfer the trust information among CSP of different DC's. The trust values of the user are computed by taking the inputs from user trust credentials and trust matrices as shown in Fig. 2. There are three distinct instances for inspection of trust value:

a) *Instance 1: Reliable user and Non-reliable user:* Here, CSP of each DC authenticates the request based on its frequency and user trust. If  $TT_{VAL} < T_{THRS D}$  then CSP of that particular DC drops the request and informs other DC's about the incoming threat.

b) *Instance 2: New/Unidentified user:* In case of any new or unidentified user, the CSP checks the trust value of the requesting user from its neighboring CSP. The fetched value is added to its own trust matrices.

c) *Instance 3: Malicious Attack:* CSP keeps track of the requests received at that distribution centre. If any suspicious activity is found then user's trust value is decreased otherwise value of trust increases.

4) For valid requests, CSP of that DC finally allow the subject to use the resource.

### B. Evaluation Startegy

The evaluation strategy of this model is setup in three phases: 1) Network phase, 2) Access control strategy and 3) Trust Evaluation strategy.

In the approach, each Distribution Centre has a table that contain trust matrix of each user associated with that DC. For every attempt by the user with the cloud resource, the DC share the trust matrices of that user to its neighboring DCs. The user requests are verified and are either granted or denied to access the service.

1) *Network model:* The network consists of several Distribution Centers(DC), cloud service providers (CSP), Owners (OWN), and Users(U).

$$DC = \sum_{i=1}^{dc} DC_i \quad \forall (1 \leq i \leq dc) \quad (1)$$

$$CSP = \sum_{i=1}^{csp} CSP_i \quad \forall (1 \leq i \leq csp) \quad (2)$$

$$OWN = \sum_{i=1}^{own} OWN_i \quad \forall (1 \leq i \leq own) \quad (3)$$

$$U = \sum_{i=1}^n u_i \quad \forall (1 \leq i \leq n) \quad (4)$$

Every pair of DCs can communicate securely and has a set of disjoint users U which means no two users can belong to same DC,  $i, j \in DC$  and therefore is written as  $U_i \cap U_j = \emptyset$ . The OWN of the resource decides the attribute values of the file, that carries a set of rights to perform a specific task on the resource.

2) *Access control strategy:* Each file stored has a set of attributes and are named as permissions in this work. The permission values to the file are decided by the owner of the file. The permission ID (PID) uniquely determines the permissions associated with that. Every Distribution Centre in the network assigns the PID to the user according to the service plan.

This approach uses the access policy similar to [31]. Here, the file/resource permissions are arranged in tabular structure with unique permission ID (PID), and its adjacent columns are represented in Boolean format. The permissions for each file are arranged as shown in the Table II.

TABLE II. PERMISSION TABLE OF THE RESOURCE

PID	Read	Write	Edit	Move	Share	Download	Delete
1	1	0	0	0	1	1	1
2	1	1	1	0	0	1	0
3	1	0	0	0	1	1	0
4	1	0	0	1	1	1	0
5	1	0	0	0	0	0	0

Each user is assigned with a unique code to access the resource. Each file will set access limits to requesting user according to permissions from set, and is represented in the equation below:

$$M \rightarrow N \quad (5)$$

Where:

"M" is a unique code in the permission table,

"N" is a set of permissions associated with Z.

For e.g., if any user has assigned with PID-4 of that file then no other user can have PID-4 for that resource/file.

3) *Trust assessment model:* The authorization process of this scheme is depended on trust-based evaluation for an efficient decision making. The users trust value is calculated and is compared with the threshold trust  $T_{THRS D}$ . After each transaction these trust values gets updated in the database. The value of the trust is related to the CSP of that particular DC in a particular session. For the evaluation of trust, the following parameters has been considered: 1) User interactions with the system  $C_a$ , 2) Type of user C, 3) The number of times the user requested for the resource  $RQ_f$ , 4) Degree of valid request  $RQ_s$  and 5) Estimated computing power of the resource  $E_{CP}$ .

The Trust matrix is of form  $\langle T_{VAL}, E_{CP} \rangle$  where:

The trust value  $T_{VAL}$  of each user for a particular session is calculated from the equations below:

$$RQ = r \times \frac{RQ_s}{RQ_f} \quad (6)$$

$$RC_{VAL} = C_t \times C_a \times RQ \quad (7)$$

Where:

$C_t$  value increases with valid requests and declines at invalid ones.

$C_a$  value greater than 0.5 is a trusted user and less than 0.5 is invalid or malicious user and equals 0.5 is for new user.

RQ value for degree of request potentiality.

$RC_{VAL}$  handles the type of user and potentiality of the request.

The ongoing limit of a cloud asset influences the exhibition of the cloud supplier and exchange execution. In this manner, the ongoing asset limit boundaries like CPU, RAM, and Network ought to be thought about while assessing trust an incentive for a cloud asset to empower a framework to gauge on the off chance that the asset can execute the expected work or not. Hence, the CPU time of resource is considered as Estimated Computing Power ( $E_{CP}$ ).

$$E_{CP} = \frac{CPU_{JOB}}{CPU_{RESOURCE}} \quad (8)$$

In this work, the trust value  $T_{VAL}$  is calculated from two attributes, the first attribute handles the frequency of the request and type of user. The second attribute handles the computing power of the resource.

$$T_{VAL} = RC_{VAL} + E_{CP} \quad (9)$$

At the end of each transaction from the DC to USER, the trust value  $T_{VAL}$  gets updated and broadcasted to the other DC's. The other DC's keeps a record of these values so that there should not be any discrepancies while handling the requests.

Algorithm: Proposed DFGACT

1. **Input:** Parameter associated with calculation of trust value
2. **Output:** Updated Trust Value
3. While(true)
4. User send request ( $T_{VAL}$ ,  $E_{CP}$ ) to DC
5. if ( $C_t \leq 1$  &&  $C_a > 0.5$ ) then
6. if ( $T_{VAL} \leq T_{THRS}$ ) then
7. Continue with the service;
8. Calculate RQ using eq.6
9. Calculate  $RC_{VAL}$  using eq.7
10. else
11.  $RQ = RQ - 1$  ;
12.  $C_t = C_t - 1$ ;
13. end if;
14. Calculate  $T_{VAL}$  from eq .9 and update the database with this new value
15. end if;
16. Broadcast ( $T_{VAL}$ ,  $E_{CP}$ ) to other DC's
17. DC extracts access level permission for the user.
18. if ( $R_{id}(U_{id}) \rightarrow true$  &&  $F_{id}$ ) then
19.  $R_{id}(U_{id}) \leftarrow P_{id}(F_{id})$
20. end if ;

21.  $RQ_s \leftarrow RQ_s + 1$
22.  $RQ_f \leftarrow RQ_f + 1$
23. end while;

#### IV. RESULTS AND ANALYSIS

This paper used a CloudSim3.0 [32] and experiment environment is eclipse editor. The programming language used is java to simulate the proposed method. It is assumed, each VM of the cloud as a Distribution Centre (DC). Likewise, 8 DC's are created for the evaluation of proposed scheme. The scheme evaluates the user's behaviour according to trust management module. At the beginning, trust value of registered users is defined as "0.5". Depending on the successful attempts, the trust value of user keeps changing, but the value should not exceed trust threshold. In this approach trust threshold is assigned as 1. The results of DFGACT and literature [30] are compared thorough simulation. The results are proved that DFGACT has better efficiency than literature [30].

Fig. 3 represents time cost of both the schemes in comparison with no. of users. From the result it can be seen that with the increasing number of users the time cost increases for Mehrjaj [30]. On the other hand, the DFGACT less delay is because this scheme has simple way of accessing request in fine-grained manner, whereas in Mehrjaj [30] the conversion of roles into tasks into permissions results in delay at higher cost that increases with the increasing no. of users.

Fig. 4 shows the network traffic rises with the no. of requests for both schemes. However, DFGACT has less overhead in comparison with the other approach. The DFGACT has less overhead as compared to Mehrjaj [30] because of DFGACT distributed nature. Mehrjaj [30] has higher overhead because of its centralized approach and at the same time various assignments of roles-tasks-permissions. Hence from the results, it is evident that DFGACT could nearly improve the security by minimizing the resources for computation.

Fig. 5 represents the rate of successful transaction with respect to time, here the DFGACT scheme has higher success rate than that of Mehrjaj [30]. The dynamic nature of user leads to increase or decrease of RST. Rather than the time factor, the person's conduct is considered for changes at the side of the variant of behavior functions.

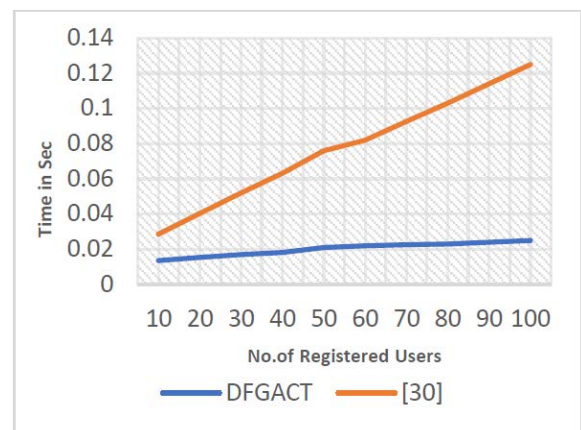




Fig. 3. Time Cost of the Schemes.

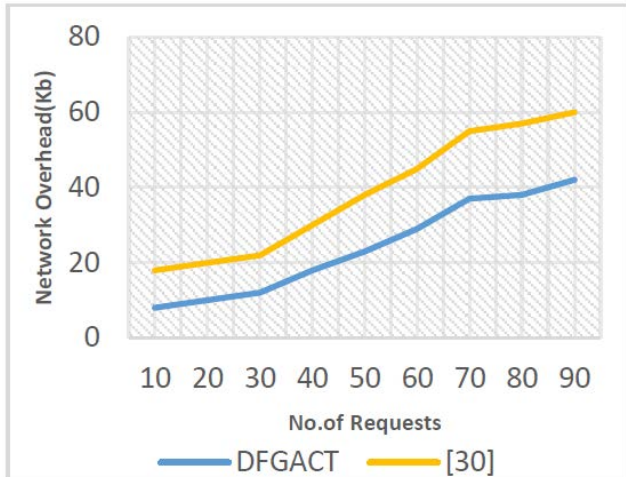


Fig. 4. Network Overhead vs. No. of Requests.



Fig. 5. Rate of Successful Transactions vs Time.

## V. CONCLUSION

In order to reduce the computational overhead due to complex equations in encryption and decryption schemes, the distributed fine-grained access control using trust assessment has been proposed. To enable cost saving, such as time cost, network overhead and rate of successful transactions the DFGACT scheme assists in creation of multiple distribution centers for providing services to the authorized users on the basis of their trust values. To achieve fine-grained access a set of access rights list is generated for each file/resource and stored in distribution centers. The evaluations had shown that this approach has 25% more better rate of successful transactions with less time and overhead than the existing ones. In future, the application of Swarm intelligence technique for the DFGACT scheme will be implemented for handling of attacks and better throughput.

## REFERENCES

- [1] Zhiguo Wan, Jun'e Liu, and Robert H. Deng (2014) "HASBE: A Hierarchical Attribute-Based Solution for Flexible and Scalable Access Control in Cloud Computing" : IEEE Transactions on Information Forensics and Security, Vol. 7, NO.2, April.
- [2] Ravi Sandhu, David Ferraiolo and Richard Kuhn (2000) "The NIST Model For Role Based access Control: Toward a Unified Standard" : ACM workshop on Role-based access control. Vol. 2000.
- [3] M. Sookhak, F. R. Yu, M. K. Khan, Y. Xiang , and R. Buyya (2017), "Attributebased data access control in mobile cloud computing: Taxonomy and open issue s," Future Gener. Comput. Syst., vol. 72, pp. 273-287.
- [4] J. Li, X. Chen, J. Li, C. Jia, J. Ma, and W. Lou (2013) "Fine-grained access control system based on outsourced attribute-based encryption," in Proc. Eur. Symp. Res. Comput. Secur., Egham, U.K.: Springer, 2013, pp. 592\_609.
- [5] K. Yang and X. Jia (2014), "Expressive, ef\_cient, and revocable data access control for multi-authority cloud storage," IEEE Trans. Parallel Distrib.Syst., vol. 25, no. 7, pp. 1735\_1744.
- [6] M. Chase and S. S. M. Chow (2009), "Improving privacy and security in multi-authority attribute-based encryption," in Proc. 16th ACM Conf. Comput. Commun. Secur. - CCS, 2009, pp. 121\_130.
- [7] S. Fugkeaw and H. Sato (2015), "An extended CP-ABE based access control model for data outsourced in the cloud," in Proc. IEEE 39th Annu. Comput. Softw. Appl. Conf., Jul. 2015, pp. 73\_78.
- [8] Agrawal, N., Tapaswi, S. (2019): A trustworthy agent-based encrypted access control method for mobile Cloud computing environment. Pervasive Mob. Comput. 52, 13–28. <https://doi.org/10.1016/j.pmcj.2018.11.003>.
- [9] Li, X., Zhou, F., Yang, X (2011):. A multi-dimensional trust evaluation model for large-scale P2P computing. J. Parallel Distrib. Comput.71(6), 837–847. <https://doi.org/10.1016/j.jpdc.2011.01.007>.
- [10] G. Lin, D. Wang, Y. Bie and M. Lei (2014), "MTBAC: A mutual trust-based access control model in Cloud computing," in China Communications, vol. 11, no. 4, pp. 154-162, April 2014, doi: 10.1109/CC.2014.6827577.
- [11] Khilar, P., Chaudhari, V., Swain, R (2019):. Trust-based access control in Cloud computing using machine learning. In: Das, H., Barik, R., Dubey, H., Roy, D. (eds) Cloud Computing for Geospatial Big Data Analytics, vol 49, pp. 55–79. Springer (2019). [https://doi.org/https://doi.org/10.1007/978-3-030-03359-0\\_3](https://doi.org/https://doi.org/10.1007/978-3-030-03359-0_3).
- [12] M. Rafiqul Islam and M. Habiba (2012), "Collaborative swarm intelligence based Trusted Computing," 2012 International Conference on Informatics, Electronics & Vision (ICIEV), 2012, pp. 1-6, doi: 10.1109/ICIEV.2012.6317341.
- [13] V. Goyal, O. Pandey, A. Sahai, and B.Waters (2006), "Attribute-based encryption for \_ne-grained access control of encrypted data," in Proc. 13th ACMConf. Comput. Commun. Secur., Alexandria, VA, USA, 2006, pp. 89\_98.
- [14] J. Bethencourt, A. Sahai, and B. Waters (2007), "C iphertext-Policy AttributeBased Encryption," in 2007 IEEE Symposium on Security and Privacy(SP '07), Berkeley, France, 2007.
- [15] S. Wang, H. Wang, J. Li, H. Wang, J. Chaudhry, M. Alazab, and H. Song (2020), "A fast CP-ABE system for cyber-physical security and privacy in mobile healthcare network," IEEE Trans. Ind. Appl., vol. 56, no. 4, pp. 4467\_4477, Jul./Aug. 2020.
- [16] K. Liang, J. K. Liu, D. S. Wong, and W. Susilo (2014), "An ef\_cient cloud-based revocable identity-based proxy re-encryption scheme for publicclouds data sharing," in Proc. Eur. Symp. Res. Comput. Secur. (EROSICS),Wroclaw, Poland, 2014, pp. 257\_272.
- [17] Ye, J., Xu, Z., Ding, Y.(2016): "Secure outsourcing of modular exponentiations in cloud and cluster computing. Clust. Comput. "19(2),811–820 (2016).
- [18] L. Zhou, V. Varadharajan, and M. Hitchens (2013), "Achieving secure role-based access control on encrypted data in cloud storage," IEEE Trans. Inf. Forensics Security, vol. 8, no. 12, pp. 1947\_1960, Dec. 2013.
- [19] Qi Han, Kan Yang, Kan Zheng, Hui Li, Xuemin Shen, Zhou Su, (2017) "An Efficient and Fine-Grained Big Data Access Control Scheme With Privacy-Preserving Policy", IEEE Internet of Things Journal, Volume:4, Issue:2, 2017.



- [20] S. Fugkeaw (2019), "A Fine-Grained and Lightweight Data Access Control Model for Mobile Cloud Computing," in *IEEE Access*, vol. 9, pp. 836-848, 2021, doi: 10.1109/ACCESS.2020.3046869.
- [21] He, H., Zheng, Lh., Li, P. et al. An efficient attribute-based hierarchical data access control scheme in cloud computing. *Hum. Cent. Comput. Inf. Sci.* 10, 49 (2020). <https://doi.org/10.1186/s13673-020-00255-5>.
- [22] Anilkumar, C., Subramanian, S. A novel predicate based access control scheme for cloud environment using open stack swift storage. *Peer-to-Peer Netw. Appl.* 14, 2372–2384 (2021). <https://doi.org/10.1007/s12083-020-00961->.
- [23] Q. Zhang, S. Wang, D. Zhang, J. Wang and Y. Zhang, "Time and Attribute Based Dual Access Control and Data Integrity Verifiable Scheme in Cloud Computing Applications," in *IEEE Access*, vol. 7, pp. 137594-137607, 2019, doi: 10.1109/ACCESS.2019.2942649.
- [24] He, H., Zhang, J., Gu, J. et al. A fine-grained and lightweight data access control scheme for WSN-integrated cloud computing. *Cluster Comput* 20, 1457–1472 (2017). <https://doi.org/10.1007/s10586-017-0863-y>.
- [25] Han, Kan Yang, Kan Zheng, Hui Li, Xuemin Shen and Zhou Su, (2016) "An Efficient and Fine-grained Big Data Access Control Scheme with Privacy-preserving Policy", *IEEE Internet of Things Journal*, 2016.
- [26] Yinglong Zhang, Shangping Wang, Yaling Zhang (2016), "A Blockchain-Based Framework for Data Sharing with Fine-grained Access Control in Decentralized Storage Systems", *IEEE Access*, Vol 4, 2016.
- [27] Li, Q., Zhu, H., Xiong, J. et al. (2019) Fine-grained multi-authority access control in IoT-enabled mHealth. *Ann. Telecommun.* 74, 389–400. <https://doi.org/10.1007/s12243-018-00702-6>.
- [28] X. Wang, L. Wang, Y. Li and K. Gai (2018), "Privacy-Aware Efficient Fine-Grained Data Access Control in Internet of Medical Things Based Fog Computing," in *IEEE Access*, vol. 6, pp. 47657-47665, 2018, doi: 10.1109/ACCESS.2018.2856896.
- [29] S. Roy, A. K. Das, S. Chatterjee, N. Kumar, S. Chattopadhyay and J. J. P. C. Rodrigues (2019), "Provably Secure Fine-Grained Data Access Control Over Multiple Cloud Servers in Mobile Cloud Computing Based Healthcare Applications," in *IEEE Transactions on Industrial Informatics*, vol. 15, no. 1, pp. 457-468, Jan. 2019, doi: 10.1109/TII.2018.2824815.
- [30] Mehraj, Saima & Bandy, M. Tariq. (2021). A flexible fine-grained dynamic access control approach for cloud computing environment. *Cluster Computing*. 24. 1-22. [10.1007/s10586-020-03196-x](https://doi.org/10.1007/s10586-020-03196-x).
- [31] Mehar, Deepak & Vishwakarma, Gagan & Jain, Yogendra. (2015). Modified Fine-grained Data Access Control Algorithms for File Storage Cloud. *International Journal of Computer Applications*. 116. 15-19. [10.5120/20467-2288](https://doi.org/10.5120/20467-2288).
- [32] R. N. Calheiros, R. Ranjan, A. Beloglazov, C. A. F. De Rose, and R. Buyya (2010), "CloudSim: A toolkit for modeling and simulation of cloud computing environments and evaluation of resource provisioning algorithms," *Softw. Pract. Exper.*, vol. 41, no. 1, pp. 23\_50, Aug. 2010, doi: 10.1002/spe.995.

#### AUTHORS' PROFILE



Ms. Aparna Manikonda worked in various prestigious institutions across India. Currently, she is pursuing Doctorate degree in computer science. She is having 15 years of teaching experience along with research and published various numerous publications. She has 8 International journal publications. She has presented papers in 15 International conferences and attended 12 Seminars/Workshops/FDP/QIP. She has organized 2 conferences, 4 workshops. She has published 2 patents and 4 books titled 'Internet and web Technology', 'Internet and Web Technology 2.0', 'Cloud computing for Beginners', 'Object Modelling using UML A software Perspective'. Her research areas include Cloud Computing, Ad-hoc and Sensor networks, IoT and Image Processing.



Dr. N. Nalini is a Professor in the Department of Computer Science and Engineering at Nitte Meenakshi Institute of Technology, Bangalore. She received her MS from BITS, Pilani in 1999 and PhD from Visvesvaraya Technological University in 2007. She has more than 24 years of teaching and 17 years of research experience. She has numerous international journal and conference publications to her credit, She has served as Technical Committee member and reviewer in various International conferences and has delivered technical talks at various Institutions. She Published book on "INTERNET OF THINGS: Advanced Wireless Technologies for Smart Ecosystems" having ISBN-13: 979-8679630055. Recognized as FSIESRP (Educational- Professional Membership Grade: Fellowship) & Editorial Board Member Registration as Hon. Consulting Editor. Program Committee Member in BIOMA International Conference since 2010. Organizing Committee Member in the Congress of the "International Conference on Electronics & Electrical Engineering" Seoul, South Korea. She has guided EIGHT candidates to complete their Ph.D successfully and currently guiding five more candidates. She has guided more than fifty PG students to complete their thesis. Her areas of research include Cryptography and Network Security, Cloud Computing, Artificial Intelligence and Machine Learning, Wireless and Distributed Sensor Networks, Optimization Heuristic Techniques.

# BERT-Based Hybrid RNN Model for Multi-class Text Classification to Study the Effect of Pre-trained Word Embeddings

Shreyashree S, Pramod Sunagar, S Rajarajeswari, Anita Kanavalli

Department of Computer Science & Engineering  
M S Ramaiah Institute of Technology (Affiliated to VTU)  
Bengaluru, India

**Abstract**—Due to the Covid-19 pandemic which started in the year 2020, many nations had imposed lockdown to curb the spread of this virus. People have been sharing their experiences and perspectives on social media on the lockdown situation. This has given rise to increased number of tweets or posts on social media. Multi-class text classification, a method of classifying a text into one of the pre-defined categories, is one of the effective ways to analyze such data that is implemented in this paper. A Covid-19 dataset is used in this work consisting of fifteen pre-defined categories. This paper presents a multi-layered hybrid model, LSTM followed by GRU, to integrate the benefits of both the techniques. The advantages of word embeddings techniques like GloVe and BERT have been implemented and found that, for three epochs, the transfer learning based pre-trained BERT-hybrid model performs one percent better than GloVe-hybrid model but the state-of-the-art, fine-tuned BERT-base model outperforms the BERT-hybrid model by three percent, in terms of validation loss. It is expected that, over a larger number of epochs, the hybrid model might outperform the fine-tuned model.

**Keywords**—Multi-class text classification; transfer learning; pre-training; word embeddings; GloVe; bidirectional encoder representations from transformers; long short-term memory; gated recurrent units; hybrid model; RNN

## I. INTRODUCTION

On March 11<sup>th</sup>, 2020, the World Health Organization (WHO) proclaimed Covid-19 a global pandemic, making human lives increasingly digital. This massive amount of digital data aids data scientists in discovering new patterns and gaining a new perspective on any area of interest. With the rise of Artificial Intelligence (AI) in data science, machines might have the ability to perform all human tasks much better than humans. Natural Language Processing (NLP), a sub-domain of AI, is an interesting research field, in which, Text Classification (TC) is a simple and yet a challenging problem that is well-recognized in the domain. It is a process of categorizing samples of text into few pre-defined categories/classes, which are of two types, viz, binary classification and multi-class text classification (MTC). Applications of TC range from sentiment analysis to topic labelling. Using TC, we can easily categorize emails, social media posts like Tweets etc. to maintain and understand the text better for making any data-driven business decisions.

The approaches to perform TC are rule-based (uses hand-crafted rules), deep-learning techniques (uses neural networks) and hybrid methods. Out of these techniques, the most significant one is the deep learning method because they are powerful and provide good results [1-2]. And this paper concentrates on classifying a Covid-19 twitter dataset of into 15 pre-defined categories. There are two parts for TC, the first part being the feature engineering, where one of its methods called word embedding is used and the second part being the classification. The main objective of this paper is to perform a comparative study on the performance of hybrid classifiers with their respective pre-trained word embedding techniques. This project performs a comparative analysis between 1) hybrid Recurrent Neural Network (RNN) model with the help of either Global Vectors for Word Representation (GloVe) and Bidirectional Encoder Representations from Transformers (BERT) pre-trained word embeddings and 2) BERT-base model. The main reason for choosing hybrid architecture over others is that it helps in boosting the performance of the overall model. With regards to embeddings methods, BERT was mainly chosen because of the following reasons.

- 1) It provides contextual embeddings.
- 2) It considers the order of words before providing the embeddings.
- 3) While other pre-trained embedding models have pre-generated embeddings, BERT has to be trained to generate dynamic embeddings (as it considers context).
- 4) It generates embeddings for Out-Of-Vocabulary words.

In order to study them, a Covid-19 twitter dataset has been used, that contains approximately two lakhs of tweets and their respective labels. There are 15 different categories of tweets in this dataset.

## II. LITERATURE SURVEY

Shah et al. [3] have developed a text classification system for BBC news by using three traditional machine learning algorithms separately, namely Logistic Regression, K- Nearest Neighbor and Random Forest Algorithms, and compared these models to choose the best one. The classification is divided into four parts, viz text pre-processing, text representation, implementation of classifier and finally the classification of news. Text pre-processing involved removing of stop words and stemming, text representation involved the use of TF-IDF

algorithm to convert the text into suitable format. The comparison between the classifiers has been done in terms of five metrics: Precision, Accuracy, F1-score, Confusion matrix and support. According to the experiment, logistic regression performed the best with 97% accuracy. Kumar et al. [4] have provided the method of text mining using popular machine learning classification algorithms and has also provided a SWOT analysis of these algorithms to summarize the work done so far in the usage of ML classification algorithms on the task of sentiment analysis, one of the major tasks of text classification. Authors have also observed that most of the classification algorithms use bag-of-words for representing text. As sentiment analysis is a significant part of text classification, it can be performed either using machine learning approach or a lexicon-based approach. Harjule et al. [5] have explored both the methods by using SentiWordNet and Word Sense Disambiguation in the former approach and Multinomial Naïve Bayes (MNB), Logistic Regression, SVM and RNN in the latter approach. Along with the above-mentioned classifiers, an ensemble classifier consisting of MNB, SVM and LR is also implemented. Later, these classifiers are being compared. The text pre-processing is done using NLTK that involves casing, removal of stop words, punctuations, URLs and hashtags, POS tagging and tokenization. The datasets used are “Sentiment140” and “Crowdfunder’s Data for Everyone library”. The observations indicate that the RNN model (LSTM) provides better results. Xia Sun et al. [6] have proposed a different approach to SA, where in the context of the text were captured using Bi-GRU and many DL models were used to classify. Among them, CNN+LSTM model outperformed all. The main focus of their work is the discovery of Drop Loss, which focuses on hard examples i.e., texts that are easier to get misclassified. This way, the classification accuracy was improved upon four sentiment datasets viz. MOOC, IMDB-2, IMDB-10 and SST-5. The CNN+LSTM architecture was also used by Giannopoulou et al. [7] to categorize e-books into pre-defined book categories using their table of contents as the text samples. Software As-A Service (SaaS) is one of the popular software delivery models. Customers should be clarified on which SaaS provider is best suited for them, in order to use cloud services. There are many service quality pillars [8] that are to be considered before choosing a right provider. Hence, Raza et al. have performed multi-class text classification on these customer reviews, with service quality pillars as categories. The classification algorithms used are machine learning algorithms and an ensemble of all the ML algorithms. The representation of text is done using TF-IDF technique after text cleaning. The results show that Logistic Regression performs better than all the models, including the ensemble model. There is also the field of citation intent classification that can be benefitted by different word embedding techniques.

Roman et al. [9] have used word embedding techniques like GloVe, InferSent and BERT to classify the citation context with citation intent, on 10 million records of Citation Context Dataset. It has been observed that the method using BERT provides a highest of 89% precision of all. Before BERT or transformer models were discovered, Bi-LSTM architectures were leading in many of the downstream tasks of

NLP. Hence, Huang et al. [10] have experimented by combining Bi-LSTM with transformers. Considering the fact that adding more hidden layers to BERT will not improve its performance, the authors have added a Bi-LSTM layer to each of the transformer entity, called TRANS-BLSTM, and have observed that their model provides an F1-score of 94.01% on SQUAD 1.1 development dataset. Hao Wu et al. [11] have proposed a weighted multi-class text classification model where the text is converted to its numerical terms using Word2Vec technique; weights are applied to those vectors using TF-IDF algorithm, and the word vectors are multiplied with these weights to provide the final representation of text. Context is captured using a BiLSTM layer, followed by an Attention layer and a softmax layer to classify. This model has observed to have 91.26% accuracy. Kumar et al. [12] have also used Bi-LSTM layers in their proposed model, SAB-LSTM, where they have applied model and network optimizer with a dropout layer around the Bi-LSTM layer to provide best accuracy when trained on COVID-19 dataset, in comparison with LSTM and Bi-LSTM models individually. Another hybrid model CNN+RNN with attention mechanism was proposed by Guo et al. [13] to perform MTC.

The text classification method also finds its application in the field of medicine prescription, where, it can be detected whether the prescribed medicine has been misused or not. Al-Garadi et al. [14] have experimented in this domain on a Twitter dataset using BERT and its variants but with fusion. The fusion models involved combining the probabilities of each text sample from BERT and its variants using either a logistic regression classifier or a Naïve Bayes classifier. These fusion models were observed to provide higher accuracy than the individual transformer models. Shaik et al. [15] have developed a text classification model that classifies the course learning outcomes (CLOs) and assessment texts into a pre-defined class of Bloom’s taxonomy, contributing to the education domain. This model uses Skip-gram word embedding technique and LSTM classifier to perform MTC, which provides an accuracy of 87% on CLOs and 74% on assessment texts. The Skip-gram technique is also used by Aslam et al. [16] to perform MTC on Google Play app reviews using CNN as its classifier with a precision of 95.49%. The CNN is also combined with Bi-LSTM using attention layer having Word2Vec as word embedding technique, proposed by Zhenget al. [17]. Similarly, CNN is combined with GRU layers as an ensemble model to perform MTC on news sources by John et al. [18], to help women select the state they want to travel or relocate to, based on the recent criminal activities. As a better alternative to CNN, CapsNets are used along with Bi-GRU layers as a hybrid model, using Word2Vec technique to perform Text classification by Gangwar et al. [19]. The detection of fake news also is a significant sub-task of MTC, where in IulianIlie et al. [20] have proposed a comparative study of 10 DNN models using GloVe, Word2Vec and FastText word embedding techniques, in which RCNN performed the best. The fastText method is used as a feature extraction method and also as a classification method to classify emails into multiple classes [21]. Aydoğan et al. [22] have performed a comparative study between CNN, LSTM, RNN and GRU networks on Turkish datasets, using word2Vec word embeddings. The results indicate that both

LSTM and GRU perform the best. Sunagar et al. [23] have conducted a comparative study between various deep learning models for the task of MTC on Covid-19 dataset, amongst which, RNN with Bidirectional LSTM performed better. The comparison between ML algorithms was also considered for the task of news topic classification [24] to study the different ML models. Many researchers have also carried out the works like detecting the disease, predicting the end of pandemic [25] and creating a decision support system [26] for Covid-19. Many authors have carried out the research on the features extraction from text and tried to establish how this will help in attaining the better accuracy [27-29].

In the existing system, classification of the text focuses on Sentiment Analysis, Movie Review etc. Due to Covid-19 pandemic, lot of tweets are being generated on various topics like, safety measures, social distancing, advisories, etc. by Government agencies, WHO, Scientists, NGOs and individuals. Classifying these tweets into different categories like Social Distancing, Vaccination, Advisories etc. is one of the motivations for the taking up this project. The recent works do contribute to the accuracy of the models discovered, may it be hybrid, traditional or an ensemble model. But this paper mainly focuses on the fact that, more the number of neural network layers, more the accuracy, with regards to the hybrid RNN model, that is inspired by the work of Sunagar et al. [30]. To boost the accuracy of the model, the BERT embeddings are being considered along with the hybrid RNN model, as they are contextual in nature and support Out-Of-Vocabulary words. And this model is being compared with the fundamental BERT-base model, by considering the importance of pre-trained word embeddings.

### III. PROPOSED MODEL

Text classification is a challenging yet interesting problem of NLP. It is a method of classifying sample texts into few pre-defined categories. The categories can be two or more in number. If the number of categories is two in number, it is called binary classification. The applications of binary classification are Sentiment Analysis, Spam filtering, Credit Card fraud detection etc. If the number of categories is more than two, then it is called multi-class text classification. The applications of MTC are Product categorization, News categorization, Citation intent classification, E-book classification etc. This project focuses on performing MTC on a COVID dataset consisting of tweets collected from Twitter and Kaggle with 15 unique categories.

#### A. Architecture of Proposed and Related Models

1) *The process of building an MTC model involves two parts:* Extraction of Embeddings: In order to perform MTC on the above dataset, the model that is built to do that only understands numerical data and not the raw text. Hence, it is necessary to convert the text samples into numerical vectors. This conversion of raw text data into numerical values is called a word embedding technique [31]. There are two types of word embedding techniques: Frequency-based and Prediction-based.

One of the earliest prediction-based embedding techniques is Word2Vec [32], which is a contextual word embedding method that provides an association between words having similar meaning. There are two models of Word2Vec method to use, in order to create word embeddings. 1) CBOW (Continuous Bag-Of-Words) model – which takes context words as input and tries to predict the target word as output. 2) Skip-gram model – predicts context words given the target word. The disadvantages of Word2Vec model are that it cannot handle out-of-vocabulary words, it relies on local information, requires large corpus to get an optimal solution and word sense is not captured separately. In order to consider the co-occurrence of words in the document, GloVe was invented [33]. This embedding technique was developed by Stanford University that is used to generate embeddings using an unsupervised approach. The training of GloVe model involves the use of global word-word co-occurrence matrix that is points out the number of times each word co-occurs with another.

Fig.1 shows an example of a co-occurrence matrix, where each row consists of unique words in the document and each column denotes the context. Here, the context length is one. E.g., It says that the word “digital” co-occurs with the word “computer” 1670 times for the selected corpus. This large matrix is factorized to provide a lower-dimensional matrix, where, each row of the lower-dimensional matrix acts as word vectors for the respective word. The model has been pre-trained to provide various dimensional word vectors. This project concentrates on using word vectors of dimension 50, where the model is trained on 6 billion tokens taken from both Wikipedia 2014 and Gigaword 5. Hence, the pre-trained word vectors will be inside a text file of name “GloVe.6B.50d.txt”. Unlike Word2Vec model, this method uses global information to construct the embeddings. The main disadvantages of using GloVe model are that it requires large memory to store the co-occurrence matrix, it cannot handle out-of-vocabulary words and word Sense is not supported. These and many other pre-trained models have one disadvantage in common, which is the unidirectional that restricts the power of those models. This led to the discovery of a specific type of transformer networks [34], BERT [35-37]. It is a pre-trained model that considers the left and right context of a word in all layers, while generating embeddings.

In order to pre-train BERT, WordPiece embeddings are used with 30,000 vocabulary size. There are two special tokens inserted into each sentence of BERT’s input, they are [CLS] and [SEP] tokens. [CLS] token represents the beginning of every sequence, which is also a classification token, for the task of NSP. Every sequence in input is separated by [SEP] token. The BERT has a powerful input representation, shown in Fig. 2, which is a combination of three embeddings: Token embeddings: Embeddings that represent each token in the document, Segment embeddings: Embeddings that are used to identify to which segment/sentence does the token belong to and Position embeddings: Embeddings representing the position of each token. There are two pre-training tasks of BERT, shown in Fig. 3, Masked Language Modeling (MLM) and Next Sequence Prediction (NSP).

	aardvark	...	computer	data	result	pie	sugar	...
cherry	0	...	2	8	9	442	25	...
strawberry	0	...	0	0	1	60	19	...
digital	0	...	1670	1683	85	5	4	...
information	0	...	3325	3982	378	5	13	...

Fig. 1. A Sample Co-occurrence Matrix.

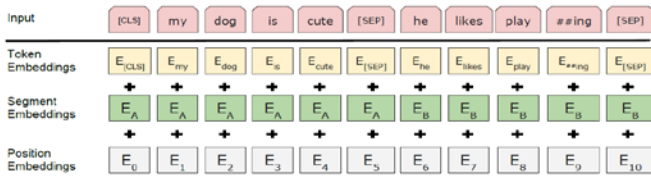


Fig. 2. Input Representation of BERT.

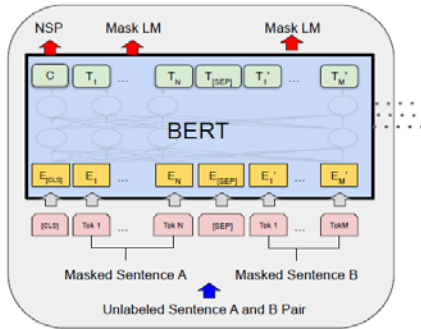


Fig. 3. Pre-training Architecture of BERT.

This paper intends to implement two word embedding methods, 1) GloVe 2) BERT-base architecture, that contains 12 transformer encoder layers, 12 self-attention heads, 768 as hidden size and 110M parameters. GloVe and BERT models are used widely due to the use of transfer learning. It is a method that saves training time of deep learning models for the data scientists. In traditional learning, the model used to be built from scratch [38] [39]. If there is a task of classification of reviews into positive and negative, and a task of classifying spam email, both these tasks were implemented separately by building two independent models from scratch. This might consume a lot of training time in general. But in case of transfer learning, the model built for classifying reviews is used as an initial checkpoint for the task of classifying spam emails. The latter task is implemented by just fine-tuning or adjusting the weights of the former model. The model that is trained for the former task is called the pre-trained model [40-41]. The latter model is called the fine-tuned model. Hence, transfer learning is a method of using the knowledge, gathered while training a model for a task, to train a similar task. Another main advantage of using transfer learning approach is that for the fine-tuning approach, the second similar task need not have a large dataset. This project uses two pre-trained models for generating word embeddings.

2) *Classifier*: At the early days of neural networks, feed forward networks (FFN) were very popular to perform many tasks. They were known for their accuracy and speed. But they had their own disadvantages:

- a) Unable to process sequential data.
- b) Current input can't be considered.
- c) Cannot remember previous inputs.

RNN (Recurrent Neural Networks) were discovered to overcome the above-mentioned problems. They belong to a class of neural networks which takes previous layer outputs and feeds them as input to the current layer, passing information from the past. This is implemented with the concept of “memory” that keeps information about previous calculations till time step  $t$ . The main reason why RNN is used in NLP is because text is a sequential data. But RNNs suffer from vanishing gradient problem, in which, the gradients are so small that the updates of parameters are insignificant. This problem occurs while processing long sequences. Hence, there are two variants of RNN that have been discovered: LSTM (Long-Short Term Memory) and GRU (Gated Recurrent Units). In RNN, in order to add new information, the whole memory context is modified and there is no consideration for important information. To overcome this, LSTM is used [42]. LSTM has the ability to forget or restore information of choice. And it is implemented by three Gating mechanisms. 1) Forget Gate: This gate is used to forget all the insignificant information from the memory context. 2) Input Gate: This gate is used to add or update new information into the memory context. 3) Output Gate: This gate is responsible for selecting important information and passing it out to the downstream network. GRU, on the other hand, is another variant of RNN similar to LSTM, except that it has two Gating mechanisms [43]: 1) Reset Gate: This gate is responsible for deciding how much of previous information should be forgotten. 2) Update Gate: This gate is responsible for deciding how much of previous information should be passed along the network. This paper uses both LSTM and GRU layers to classify tweets. Fig. 4 shows the architecture of the proposed hybrid RNN model.

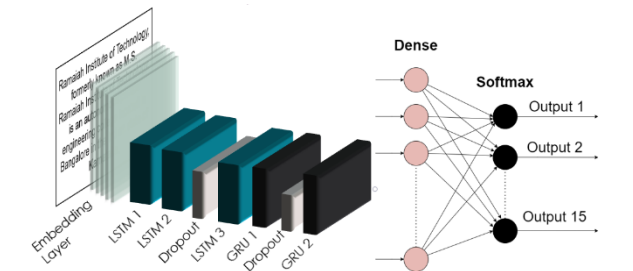


Fig. 4. Architecture Diagram of Hybrid Model.

The input word sequence is provided to the input layer, where tokenization (Word/Sub-word, depending on the embedding technique) takes place and the tokenized data is given to the second layer (Embedding Layer). This layer can use either GloVe or BERT pre-trained word embedding techniques to generate meaningful word embeddings. These embeddings are given to the third layer, which is a combination of LSTM, Dropout and GRU layers. This is the hybrid classifier. The fourth layer is the softmax output layer which gives out the probability for all the classes, with highest probability for the predicted class.

### B. Working of Models

Fig. 5 depicts the generic workflow. The dataset used in this project contains tweets collected from Twitter till April 2020. The total number of tweets collected are 260000. There are three attributes in the dataset: Tweets, labels and label ids.

All the tweets are labelled as one of the 15 labels. These tweets, before feeding into the models, have to be cleaned and pre-processed such that it is easier for the models to learn quickly.

---

### Algorithm for proposed model

---

**Input:** The COVID dataset

**Output:** A model trained on the COVID dataset and one of the 15 pre-defined classes for each tweet in the test dataset

---

1. Import the dataset.
  2. Pre-processed\_tweets, labels = Data\_PreProcessing(dataset).
  3. Split the Pre-processed\_tweets and labels into training and testing set with a ratio of 80:20.
  4. Either perform BERT\_Tokenization() or GloVe\_Tokenization() depending on the choice of embedding technique.
  5. Create an embedding matrix for every word in the vocabulary
  6. Build hybrid model as shown in step 7 to step 13.
  7. Add an EmbeddingLayer() with weights as the embedding matrix
  8. 3 LSTM() Layer
  9. Dropout layer
  10. 1 LSTM() layer& 1 GRU Layer
  11. Dropout layer
  12. 1 GRU() Layer
  13. Dense() layer with Softmax activation function.
  14. Train the model on the training set.
  15. Evaluate the model on the test set.
- 

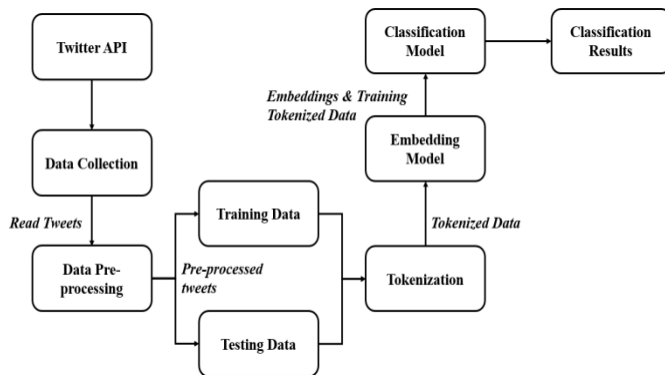


Fig. 5. Generic Workflow.

After the pre-processing, the dataset is split into training and testing set with 80:20 ratio, where each sample from each set is tokenized (GloVe or BERT Tokenization) and is given to the embedding model (GloVe or BERT) to generate the embeddings which is given to the classification model (Hybrid or BERT-base) to produce the results (one of the 15 classes of the dataset). There are two ways to use BERT model. 1) Fine-tuning approach [44] – Using the BERT model as a whole to generate embeddings from 12 encoders and 1 classifier. 2) Embeddings extraction approach – Using BERT model to just extract the embeddings and use a different classifier altogether. This project suggests the use of Hybrid RNN model for this approach as a classifier. The reason being,

LSTM is more accurate and complex compared to GRU because of the number of gating mechanisms and in terms of speed, GRU is better than LSTM. Hence, to combine the advantages of both the models this hybrid model has been proposed. For the second approach, the embeddings can be extracted in various ways. We can extract the embeddings from 1) the last layer 2) the last few layers and sum them 3) the last few layers and get an average. This paper uses the second approach. Following is the algorithm for the approach towards the hybrid model.

## IV. RESULTS AND ANALYSIS

### A. Dataset

This paper uses a COVID-19 dataset that is prepared with tweets from twitter and Kaggle in the period 2020-21. The structure of this dataset contains tweets, labels in words and label ids. In terms of MTC, the dataset contains 15 unique classes to categorize the tweets that are: Entertainment, Essential Workers, Facts, General, Government Action, Medical Test & Analysis/Supply, Tribute, Pandemic, Panic Shopping, Political, Self-Care, Social Distance, Stay-At-Home, Taco Tuesday and Telecommuting Life.

### B. Experimental Settings

The setup requires Google Colab Pro subscription that is linked to Google Drive, where the dataset is contained. The BERT repository is cloned in the Colab platform, and is used to access files that help in extracting embeddings from BERT's encoder layers. The python files extracted for this purpose are modified for the current use case accordingly. Also, the configuration files of uncased-BERT model, "uncased\_L-12\_H-768\_A-12", is downloaded to the drive, that contains vocabulary file, configuration file and checkpoint of pre-trained BERT-base model. These files are given as parameters to the BERT repository files to extract embeddings. Following is the list of all main parameters considered for configuring all 3 models:

For both the Hybrid models:

- Number of LSTM layers: 3
- Number of GRU layers: 2
- Number of dropout layers: 2
- Output function: SoftMax
- Learning rate: 0.001
- Batch size: 256
- Optimizer: Adam
- Dropout rate: 0.5

For BERT-Hybrid model:

- Maximum sequence length: Maximum sentence length
- Embedding dimension: 768
- Vocabulary size: 30522
- No. of encoder layers for embeddings extraction: 4



For GloVe-Hybrid model:

- Maximum sequence length: 500
- Embedding dimension: 256

Vocabulary size: 1 lakh approx.

### C. Performance Measures

Table I describe the validation loss, validation accuracy, precision, recall and F1-score of BERT-hybrid, GloVe-hybrid and BERT-base models. From the tables, we can say that, for three epochs, BERT-base model performs better than the hybrid models with an accuracy of 96.59%. It is estimation that for larger number of epochs, hybrid model might work better than the base model. If we compare between the two hybrid models, the model that uses BERT embeddings shows a slight improvement in performance, indicating that the use of embeddings plays a major role in deciding the performance of any model.

TABLE I. LOSS, ACCURACY, PRECISION, F1-SCORE AND RECALL OF MODELS

MODEL	LOSS	ACCURACY	PRECISION	RECALL	F1-SCORE
GLOVE-HYBRID	0.156	0.953	95.74	95	95.34
BERT-HYBRID	0.1475	0.9568	95.75	95.55	95.63
BERT-BASE	0.1185	0.9659	96.93	96.85	96.88

Fig. 6 depicts the training and prediction time of implemented models. As the BERT-base model is more complex in architecture, it takes more time to train and predict. BERT-hybrid model takes least amount of time to train, with a smaller number of parameters and small vocabulary size. Also, the time-consuming task of generating embeddings is a one-time process for the BERT-hybrid model, and hence the small training time. The GloVe hybrid model takes more time to train than BERT-hybrid model due to its larger vocabulary size and parameters. With respect to the prediction time, BERT-hybrid takes less time to predict than the other models, approximately five minutes, which leads to a fact that there is a trade-off between time and accuracy when we use BERT-hybrid model. Fig. 7 and Fig. 8 below show the comparison between models on validation accuracy and validation loss in a graphical format respectively.

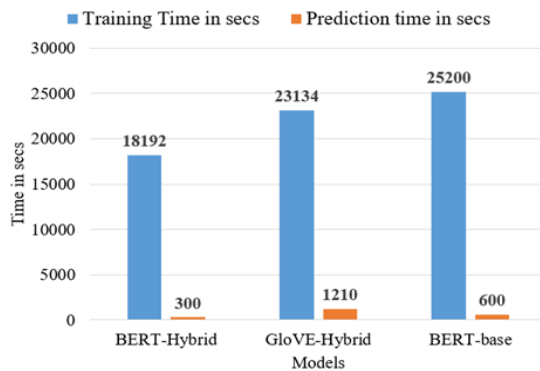


Fig. 6. Training and Prediction Time of Models.

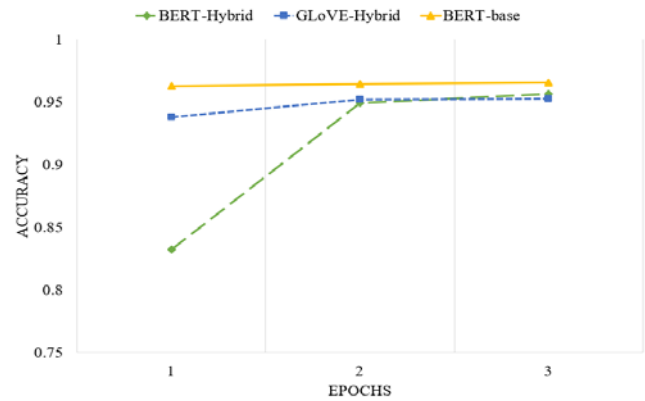


Fig. 7. Validation Accuracy of Models.

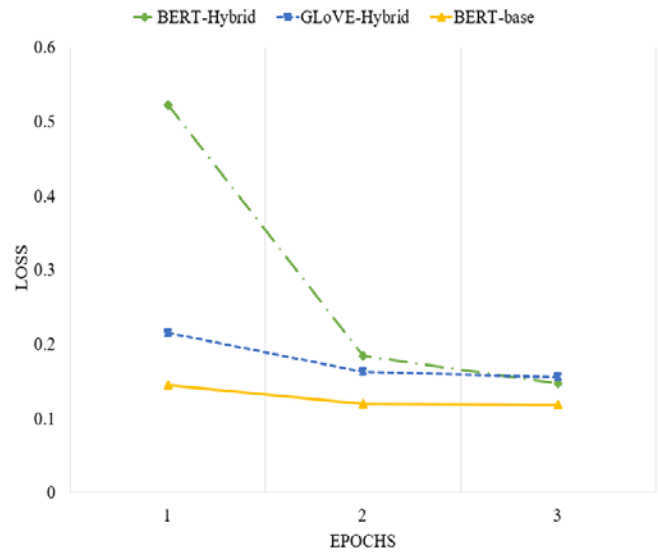


Fig. 8. Validation Loss of Models.

### V. CONCLUSION AND FUTURE WORK

The research presents a novel hybrid RNN model that experiments between GloVe and BERT embeddings. In terms of accuracy, performance, and speed, this hybrid model utilizes the capabilities of both the LSTM and GRU layers to fill in the gaps. It also uses many layers of LSTM and GRU to apply the concept that deeper the model, greater the accuracy. For three epochs, it is shown that the state-of-the-art BERT-base transformer model outperforms both hybrid RNN models, with an accuracy of 96.59 %. It is expected that the hybrid RNN models will perform better over a larger number of epochs. Furthermore, the BERT-hybrid model outperforms the GloVe-hybrid model, demonstrating that contextual representation improves performance. In the future, different BERT model versions might be utilized to produce embeddings and feed them to the hybrid model for better performance.

### ACKNOWLEDGMENT

This work was supported by M S Ramaiah Institute of Technology, Bangalore-560054, and Visvesvaraya Technological University, Jnana Sangama, Belagavi -590018.

REFERENCES

- [1] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, J. Gao, "Deep Learning--Based Text Classification: A Comprehensive Review", *ACM Computing Surveys (CSUR)*, vol. 54, pp. 1-40, 2021.
- [2] S. Dong, P. Wang, K. Abbas, "A Survey on Deep Learning and its Applications", *Computer Science Review*, vol. 40, p. 100379, 2021.
- [3] K. Shah, H. Patel, D. Sanghvi, M. Shah, "A Comparative Analysis of Logistic Regression, Random Forest and KNN Models for the Text Classification", *Augmented Human Research*, vol. 5, pp. 1-16, 2020.
- [4] A. Kumar, V. Dabas, P. Hooda, "Text Classification Algorithms for Mining unstructured data: a SWOT analysis", *International Journal of Information Technology*, vol. 12(4), pp. 1159-1169, 2020.
- [5] P. Harjule, A. Gurjar, H. Seth, P. Thakur, "Text classification on Twitter data", in *3rd International Conference on Emerging Technologies in Computer Engineering: Machine Learning and Internet of Things (ICETCE)*, pp. 160-164, IEEE, 2020.
- [6] X. Sun, Y. Gao, R. Sutcliffe, S.X. Guo, X. Wang, J. Feng, "Word Representation Learning Based on Bidirectional Grus with Drop Loss for Sentiment Classification", *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 51, pp. 4532-4542, 2021.
- [7] E. Giannopoulou, N. Mitrou, "An AI-Based Methodology for the Automatic Classification of a Multiclass Ebook Collection Using Information from the Tables of Contents", *IEEE Access*, vol. 8, pp. 218658-218675, 2020.
- [8] M. Raza, F.K. Hussain, O.K. Hussain, M. Zhao, Z. Rehman, "A Comparative Analysis of Machine Learning Models for Quality Pillar Assessment of Saas Services by Multi-Class Text Classification of Users' Reviews", *Future Generation Computer Systems* 101, pp. 341-371, 2019.
- [9] M. Roman, A. Shahid, S. Khan, A. Koubaa, L. Yu, "Citation Intent Classification Using Word Embedding", *IEEE Access*, vol. 9, pp. 9982-9995, 2021.
- [10] Z. Huang, P. Xu, D. Liang, A. Mishra, B. Xiang, "TRANS-BLSTM: Transformer with Bidirectional LSTM for Language Understanding", *arXiv preprint arXiv:2003.07000*, 2020.
- [11] H. Wu, Z. He, W. Zhang, Y. Hu, Y. Wu, Y. Yue, "Multi-Class Text Classification Model Based on Weighted Word Vector and Bilstm-Attention Optimization", in *International Conference on Intelligent Computing*, Springer, Cham., pp. 393-400, 2021.
- [12] D.A. Kumar, A. Chinnalagu, "Sentiment and Emotion in Social Media COVID-19 Conversations: SAB-LSTM Approach" in *9th International Conference System Modeling and Advancement in Research Trends (SMART)*, IEEE, pp. 463-467, 2020.
- [13] L. Guo, D. Zhang, L. Wang, H. Wang, B. Cui, "CRAN: A Hybrid CNN-RNN Attention-Based Model for Text Classification" in *International Conference on Conceptual Modeling*, pp. 571-585, Springer, Cham, 2018.
- [14] M.A Al-Garadi, Y.C Yang, H. Cai, Y. Ruan, K. O'Connor, G.H Graciela, J. Perrone, A. Sarker, "Text Classification Models for the Automatic Detection of Nonmedical Prescription Medication Use from Social Media", *BMC Medical Informatics and Decision Making*, vol. 21, pp. 1-3, 2021.
- [15] S. Shaikh, S.M. Daudpotta, A.S. Imran, "Bloom's Learning Outcomes' Automatic Classification Using Lstm and Pretrained Word Embeddings", *IEEE Access*, vol. 9, pp. 117887-117909, 2021.
- [16] N. Aslam, W.Y. Ramay, K. Xia, N. Sarwar, "Convolutional Neural Network Based Classification of App Reviews", *IEEE Access*, vol. 8, pp. 185619-185628, 2020.
- [17] J. Zheng, L. Zheng, "A Hybrid Bidirectional Recurrent Convolutional Neural Network Attention-Based Model for Text Classification", *IEEE Access*, vol. 7, pp. 106673-106685, 2019.
- [18] J. John, M.S. Varkey, M. Selvi, "Multi-Class Text Classification and Publication of Crime Data from Online News Sources" in *8th International Conference on Smart Computing and Communications (ICSCC)*, pp. 64-63, IEEE, 2021.
- [19] A.K. Gangwar, V. Ravi, "A Novel Bgcapsule Network for Text Classification", *SN Computer Science*, vol. 3, pp. 1-2, 2022.
- [20] V.I. Ilie, C.O. Truică, E.S. Apostol, A. Paschke, "Context-Aware Misinformation Detection: a Benchmark of Deep Learning Architectures Using Word Embeddings", *IEEE Access*, vol. 9, pp. 162122-162146, 2021.
- [21] R. Tahsin, M.H. Mozumder, S.A. Shahriyar, M.A. Mollah, "A Novel Approach for E-Mail Classification Using Fasttext" in *IEEE Region 10 Symposium (TENSymp)*, pp. 1392-1395, 2020.
- [22] M. Aydoğan, A. Karci, "Improving the Accuracy Using Pre-Trained Word Embeddings on Deep Neural Networks for Turkish Text Classification", *Physica A: Statistical Mechanics and its Applications*, vol. 541, p. 123288, 2020.
- [23] P. Sunagar, A. Kanavalli, V. Poornima, V.M. Hemanth, K. Sreeram, K.S. Shivakumar, "Classification of Covid-19 Tweets Using Deep Learning Techniques" in *Inventive Systems and Control*, pp. 123-136, Springer, Singapore, 2021.
- [24] P. Sunagar, A. Kanavalli, S.S. Nayak, S.R. Mahan, S. Prasad, "News Topic Classification Using Machine Learning Techniques" in *International Conference on Communication, Computing and Electronics Systems*, pp. 461-474, Springer, Singapore, 2021.
- [25] S. Shwetha, P. Sunagar, S. Rajarajeswari, A. Kanavalli, "Ensemble Model to Forecast the End of the COVID-19 Pandemic" in *3rd International Conference on Communication, Computing and Electronics Systems*, pp. 815-829, Springer, Singapore, 2022.
- [26] F. Saleem, A.S. AL-Ghamdi, M.O. Alassafi, S.A. ALGhamdi, "Machine Learning, Deep Learning, and Mathematical Models to Analyze Forecasting and Epidemiology of COVID-19: A Systematic Literature Review", *International Journal of Environmental Research and Public Health*, vol. 9, p. 5099, 2022.
- [27] Chandrika, C. P., & Kallimani, J. S. (2022). Authorship Attribution for Kannada Text Using Profile Based Approach. In *Proceedings of the 2nd International Conference on Recent Trends in Machine Learning, IoT, Smart Cities and Applications* (pp. 679-688). Springer, Singapore.
- [28] Chandrika, C. P., & Kallimani, J. S. (2022). Instance Based Authorship Attribution for Kannada Text Using Amalgamation of Character and Word N-grams Technique. In *Distributed Computing and Optimization Techniques* (pp. 547-557). Springer, Singapore.
- [29] P. Sunagar, A. Kanavalli and N D Shetty, "Feature Extraction and Selection Techniques for Text Classification: A Survey", *International Journal of Advanced Research in Engineering and Technology*, 11(12), 2020, pp. 2871-2881. doi: 10.34218/IJARET.11.12.2020.268.
- [30] P. Sunagar and A. Kanavalli, "A Hybrid RNN based Deep Learning Approach for Text Classification" *International Journal of Advanced Computer Science and Applications(IJACSA)*, 13(6), 2022. <http://dx.doi.org/10.14569/IJACSA.2022.0130636>
- [31] D. Chandrasekaran, V. Mago, "Comparative Analysis of Word Embeddings in Assessing Semantic Similarity of Complex Sentences", *IEEE Access*, vol. 9, pp. 166395-166408, 2021.
- [32] T. Mikolov, K. Chen, G. Corrado, J. Dean, "Efficient Estimation of Word Representations in Vector Space", *arXiv preprint arXiv:1301.3781*, 2013.
- [33] J. Pennington, R. Socher, C.D. Manning, "GloVe: Global Vectors for Word Representation" in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532-1543, 2014.
- [34] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, "Attention is All You Need", *Advances in neural information processing systems*, vol. 30, 2017.
- [35] J. Devlin, M.W. Chang, K. Lee, K. Toutanova, "Bert: Pre-Training of Deep Bidirectional Transformers for Language Understanding", *arXiv preprint arXiv:1810.04805*, 2018.
- [36] R.K. Kaliyar, "A Multi-Layer Bidirectional Transformer Encoder for Pre-Trained Word Embedding: A Survey of Bert" in *10th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, pp. 336-340, IEEE, 2020.
- [37] S. Shreyashree, P. Sunagar, S. Rajarajeswari, A. Kanavalli, "A Literature Review on Bidirectional Encoder Representations from Transformers", *Inventive Computation and Information Technologies*, pp. 305-320, 2022.

- [38] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, Q. He, "A Comprehensive Survey on Transfer Learning" in *Proceedings of the IEEE*, vol. 109, pp. 43-76, 2020.
- [39] H. Liang, W. Fu, F. Yi, "A Survey of Recent Advances in Transfer Learning" in *19th International Conference on Communication Technology (ICCT)*, pp. 1516-1523, IEEE, 2019.
- [40] J. Peng, K. Han, "Survey of Pre-Trained Models for Natural Language Processing" in *International Conference on Electronic Communications, Internet of Things and Big Data (ICEIB)*, pp. 277-280, IEEE, 2021.
- [41] X. Qiu, T. Sun, Y. Xu, Y. Shao, N. Dai, X. Huang, "Pre-Trained Models for Natural Language Processing: A Survey", *Science China Technological Sciences*, vol. 63, pp. 1872-1897, 2020.
- [42] R.C. Staudemeyer, E.R. Morris, "Understanding LSTM--A Tutorial into Long Short-Term Memory Recurrent Neural Networks", arXiv preprint arXiv:1909.09586, 2019.
- [43] R. Dey, F.M. Salem, "Gate-Variants of Gated Recurrent Unit (GRU) Neural Networks" in *60th International Midwest Symposium on Circuits and Systems (MWSCAS)*, pp. 1597-1600, IEEE, 2017.
- [44] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, "Huggingface's Transformers: State-of-the-Art Natural Language Processing", arXiv preprint arXiv:1910.03771, 2019.

# A Hybrid Approach of Wavelet Transform, Convolutional Neural Networks and Gated Recurrent Units for Stock Liquidity Forecasting

Mohamed Ben Houad<sup>1</sup>, Mohammed Mestari<sup>2</sup>, Khalid Bentaleb<sup>3</sup>, Adnane El Mansouri<sup>4</sup> and Salma El Aidouni<sup>5</sup>  
Laboratory of 2IACS, ENSET, Hassan II University, PO Box 159 Bd Hassan II,  
28830, Mohammedia, Morocco

**Abstract**—Stock liquidity forecasting is critical for investors, issuers, and financial market regulators. The object of this study is to propose a method capable of accurately predicting the liquidity of stocks. The few studies on stock liquidity forecasting have focused on single models such as Seasonal Auto-Regressive Integrated Moving Average with exogenous factors, the nonlinear autoregressive network with exogenous input, and Deep Learning. A new trend in forecasting which attempts to combine several approaches is emerging at the moment. Inspired by this new trend, we propose a hybrid approach of Wavelet Transform, Convolutional Neural Networks, and Gated Recurrent Units to predict stock liquidity. Our model is tested on daily data of companies listed on the Casablanca Stock Exchange from 2000 to 2021. Its forecasting performances are evaluated based on the Mean Absolute Error, the Root Mean Square Error, the Mean Absolute Percentage Error, Theil's U statistic, and the correlation coefficient. Finally, the outperformance of the proposed model is confirmed by comparison with other reference forecasting models. This study contributes to the enrichment of the field of prediction of financial risks and can constitute a framework of analysis allowing to help the stakeholders of the financial markets to forecast the liquidity of the actions.

**Keywords**—Stock liquidity; wavelet transform; convolutional neural networks; GRU cell; Casablanca stock exchange

## I. INTRODUCTION

"A stock is considered liquid if large transactions can be made rapidly without significantly impacting the stock price and without incurring substantial losses, and if any price variation caused by a random shock is quickly adjusted" [1]. Before the financial crisis of 2007-2008, stock liquidity risk was largely underestimated by investors, financial market regulators, and researchers. Yet, it has negative financial and economic consequences. In fact, It increases equity market risk [2], [3], [4], and reduces bank liquidity [5], [6]. Stock liquidity affects financial stability [7], [8]. It also impacts the financial structure and cost of capital [9], [10], the dividend distribution policy [11], and the risk of corporate failures [12], [13].

Forecasting stock liquidity is crucial for investors, issuers, and financial market regulators. It allows investors to forecast the illiquidity premium to be charged to compensate for lower returns [14]. Stock liquidity prediction helps issuers to choose the right time to go public, increase their capital or carry out financial packages such as takeover bids, sales, or the outs. Financial market regulators are also concerned with liquidity predictions as it allows them to act a priori to safeguard financial stability.

Only a few studies have attempted to address forecasting stock liquidity. We are aware of only two research articles in this area. In a comparative study, [15], concludes that the Nonlinear Autoregressive Network with exogenous inputs (NARX) has better predictive performance than the Seasonal Auto-Regressive Integrated Moving Average with exogenous factors (SARIMAX). These authors found that SARIMAX method is inaccurate because stock liquidity is irregular, noisy, and nonlinear time series. However, this study has some shortcomings.

It is based on only 108 observations, a number that we consider to be low for the learning processes of an algorithm capable of making effective predictions. The training of this type of algorithm on the basis of the large dataset can generate the problems of vanishing and exploding of the gradient. Since the number of hidden layers is low, NARX neural networks cannot capture hidden functional relationships in the historical stock liquidity dataset. As a result, their predictive performance is very limited.

The author in [16] compared linear regression model, multilayer perceptron, and Long Short Term Memory (LSTM). Based on daily data of companies listed on the Ho Chi Min and Hanoi Stock Exchanges in Vietnam from January 2011 to December 2019, the authors conclude that the LSTM model has the lowest Mean Square Error (MSE). This result seems logical to us because LSTM neural networks have more advantages than the linear regression model and the multilayer perceptron. Compared to linear regression, the LSTM model capture the characteristics of even nonlinear data. The multilayer perceptron assumes that inputs and outputs are independent of each other. On the other hand, the LSTM model takes into account the temporal dependencies while avoiding the problem of vanishing gradient. However, despite the advantages of the LSTM model, the results of this study are less convincing. First, the predictive performance is only evaluated on the basis of a single criterion (MSE). Second, despite their power, LSTM neural networks alone cannot capture all abrupt and dynamic changes in financial time series [17]. Stock liquidity is noisy, volatile, and non-linear. It requires pre-processing before forecasting.

The Wavelet Transform (WT) is an effective tool for denoising the most complex time series. By decomposing a signal into different scales, the WT can capture the hidden features of the time series. Therefore, to further improve the efficiency and accuracy of forecasting, researchers have started

to develop hybrid models that combine WT and deep learning algorithms. These hybrid models are used to predict time series, such as wind [18], solar energy [19], water quality [20], nickel futures price [21], gold returns [22], and stock prices [23]. Exploring this approach, we propose a hybrid model that combines a Wavelet Transform (WT), a Convolutional Neural Network (CNN), and a Recurrent Neural Network (RNN) with a Gated Recurrent Unit (GRU) layer. The purpose of this study is to find out if stock liquidity denoising by WT can improve the predictive performance of deep learning algorithms. The proposed WT-CNN-GRU model is tested on daily data of companies listed on the Casablanca Stock Exchange from 2000 to 2021. Our database contains 5478 trading days.

This study brings three novelties. First, the originality of our approach is the denoising of stock liquidity data by the WT before proceeding to the forecasting by deep learning algorithms. Second, unlike previous studies that used unique signal decomposition methods to denoise the data, we use adaptive approaches consistent with the characteristics of different measures of stock liquidity. Third, the proposed model showed better performance in accurately predicting the strong disruptions in stock liquidity caused by the COVID-19 pandemic.

While the predictive performance of the previously presented models is measured by a single criterion, the proposed model in this study is evaluated by a multitude of parameters such as Mean Absolute Error, the Root Mean Square Error, the Mean Absolute Percentage Error, Theil's U statistic and the correlation coefficient. By showing superiority over previous studies, the proposed model is considered a step forward in improving the prediction of stock liquidity.

The rest of the paper proceeds as follows. The next section describes the adopted methodology. Section 3 presents the empirical process, results, and comparative analysis of stock liquidity forecasting. Section 4 presents the conclusions and offers some suggestions and perspectives.

## II. PROPOSED METHODOLOGY

To forecast stock liquidity, the proposed methodology is a hybrid approach between a WT, CNN, and GRU. Fig. 1 shows the general procedure of our model.

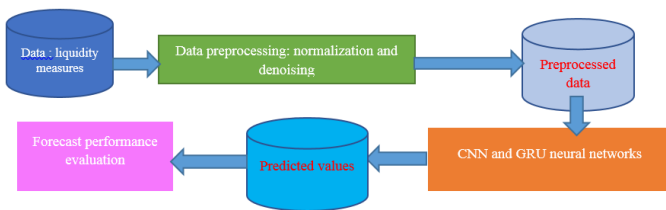


Fig. 1. Flowchart of the WT-CNN-GRU Model.

The normalized and denoised stock liquidity measures are inputs to the mixed CNN-GRU model. The detailed steps for processing the pre-processed data in the CNN-GRU model are shown in Fig. 2.

This section discusses the main steps of the hybrid WT-CNN-GRU model: (1) data preprocessing (2) a hybrid CNN-

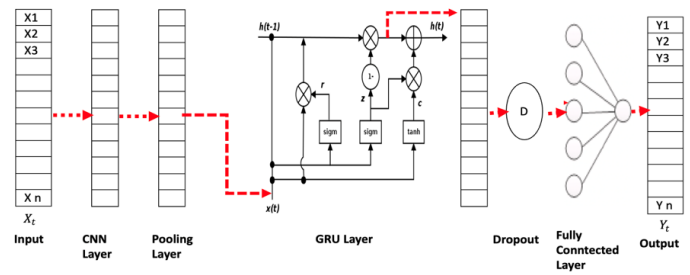


Fig. 2. Flow Chart of the CNN-GRU Model.

GRU neural network and (3) forecasting performance evaluation.

### A. Data Pre-Processing

1) *Data Normalization*: The normalized data is calculated using the Z-Score method. Z-scores measure the distance between a data point and the means in terms of standard deviation. The standardized data set has a mean of 0 and a standard deviation of 1, and retains the shape properties of the original data set (same skewness and kurtosis). The standardized data are obtained by (1).

$$x^* = \frac{(x - \bar{x})}{\sigma} \quad (1)$$

$\bar{x}$  is the mean of the original data and  $\sigma$  is the standard deviation of the original data. Normalization supports machine learning algorithms in measuring the distance between the standard deviation and the mean of processed data samples. Conversely, The original data can be derived as follows:

$$x = \sigma x^* + \bar{x} \quad (2)$$

2) *Wavelet Transform*: Financial series are noisy, volatile, nonlinear, and non-stationary. As a consequence, they require pre-processing. The Discrete Wavelet Transform (DWT) is a mathematical tool that decomposes the input signal into several physically significant components, invisible in the raw data. These components can be frequencies, trends, edges, or breaks. This facilitates the analysis of each component in isolation and the reconstruction of the original signal into the desired components to be extracted. This facilitates the denoising of the input signal. The purpose of this step in our model is to eliminate the noise that can characterize stock liquidity.

There are several signal decomposition techniques such as Maximum Overlap Discrete Wavelet Transform (MODWT), Empirical Mode Decomposition (EMD), Empirical Wavelet Transform (EWT), Tunable Q-factor Wavelet Transform (TQWT), and Variational Mode Decomposition (VMD). The choice between these techniques depends on the characteristics of the input signal [1]. MODWT is adapted for signals containing oscillations with trends or transitions. TQWT and VMD are most suitable for signals containing high or low-frequency oscillations. EWT is intended for extracting low-frequency oscillations. EMD is used when the input signal contains trends.

However, before proceeding with the decomposition of input signals, it is necessary to analyze them in time frequency to choose the most appropriate decomposition method. The Continuous Wavelet Transform (CWT) is more efficient for performing time-frequency analysis of a signal than the DWT because the scales are discretized more finely in CWT. In this study, we prefer Morse Wavelets because it has the advantage of varying amplitude and frequency over time [24].

### B. A Hybrid CNN-GRU Model

1) *Convolutional Neural Network*: A Convolutional Neural Network (CNN) is a network architecture for deep learning that learns directly from the data, eliminating the need for manual feature extraction. Fig. 3 shows a simple CNN architecture.

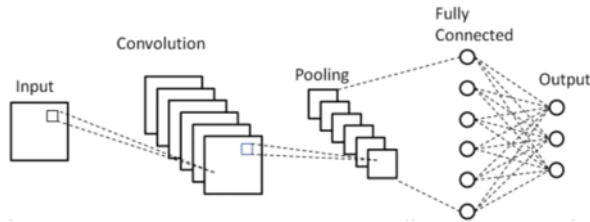


Fig. 3. Basic Architecture of the Convolutional Neural Network.

In addition to the input and output layers, three different layers are normally present in CNNs, such as the convolution layer, the pooling layer, and the fully connected layer. The convolution layer is a set of filters whose purpose is to extract local features from the input layer. This ensures that the network focuses on low-level features in the first hidden layer, then it assembles them into higher-level features in the next hidden layer, etc. Convolution layers are used in our study to extract chaotic, irregular, and fluctuating features from liquidity measurements. Pooling layers are used to retain only the most relevant features of the liquidity measures and to deepen them. Pooling can be of two types, maximum pooling, and average pooling. In this study, we retain maximum pooling because pooling by the mean is an outlier. The fully connected layer is similar to a Feedforward Neural Network whose goal is to extract the global feature of the inputs. Each neuron in these layers is connected to all hidden neurons in the previous layer.

2) *Recurrent Neural Network: The Closed Recurrent Unit Layer*: In contrast to deep Multilayer Perceptron (MLP), Recurrent Neural Networks (RNN) have interdependent input and output layers. RNNs are suitable for modeling sequential data and their associated temporal dynamics with greater accuracy. However, simple RNNs are generally characterized by the vanishing gradient problem, where, depending on the activation function, information "vanish with time," and the term nonlinearity is often inadequate for longer-term memory. To overcome this problem, Long Short-Term Memory networks (LSTMs) have been developed. LSTMs help to preserve errors that can be back-propagated across time and layers. By maintaining these errors, LSTMs allow RNNs to continue learning more efficiently across many time steps.

Compared to LSTM networks, GRUs [25], have only two gates; a reset gate and an update gate. The update gate behaves

similarly to the forget gate in LSTM by deciding which information to keep and which new information to add, while the reset gate is another mechanism to determine the amount of past temporal information to delete. We retain GRUs in our model because they are less easy to construct than LSTMs, due to fewer tensor operations.

As mentioned earlier, this study proposes a hybrid model of CNN and GRU, with corresponding parameters summarized in Table I.

TABLE I. PARAMETER OF THE CNN-GRU MODEL

Hybrid model	Parameter	Values	
Network CNN	Input layer	1	
	Dimension convolution layer	1 D	
	Number of convolution layers	2	
	Number of Pooling layers	2	
	Number of filters	1	
	Width of filters	1	
	Pooling method	Max Pooling	
	GRU network	Number of GRU layers	1
		Number of masked units	300
		Activation function to update the masked state	Tanh
Activation function to be applied to gates		Sigmoid	
Dropout layer Options	dropout rate	0.4	
Options Learning Network CNN-GRU	Epochs	300	
	optimizer	Adam	
	Initial learning rate	0.005	
	Loss function	Mean Square Error	

3) *Evaluation of the Forecasting Performance*: To evaluate the prediction performance of the proposed model, five statistical evaluation indicators are used to compare the performance of the associated models, including MAE (Mean Absolute Error), MAPE (Mean Absolute Percentage Error), RMSE (Root Mean Square Error), Theil's U statistic and the correlation coefficient ( $R^2$ ) which can be calculated as follows:

$$MAE(X, \vec{X}) = \frac{\sum_{t=1}^T \|X_t - \vec{X}_t\|}{T} \quad (3)$$

$$MAPE(X, \vec{X}) = \left( \sum_{t=1}^T \frac{\|X_t - \vec{X}_t\|}{X_t} \right) / T \quad (4)$$

$$RMSE(X, \vec{X}) = \sqrt{\frac{\sum_{t=1}^T (X_t - \vec{X}_t)^2}{T}} \quad (5)$$

$$U(X, \vec{X}) = \left( \sqrt{\frac{\sum_{t=1}^T (X_t - \vec{X}_t)^2}{T}} \right) / \left( \sqrt{\frac{\sum_{t=1}^T (\vec{X}_t)^2}{T}} + \sqrt{\frac{\sum_{t=1}^T (X_t)^2}{T}} \right) \quad (6)$$



$$R^2(X.\vec{X}) = 1 - \left( \frac{\sum_{t=1}^T (X_t - \vec{X}_t)^2}{\sum_{t=1}^T (X_t - \bar{X})^2} \right) \quad (7)$$

Where  $T$  is the number of observations,  $X$  is the actual value,  $\vec{X}$  is the forecasted value.

### III. EMPIRICAL RESULTS AND DISCUSSION

#### A. Sample, Liquidity Measures, and Data Analysis

1) *Sample and Liquidity Measures:* To test our model, we collect the best bid and ask price, closing price, and daily trading volume of 75 companies listed on the C.S.E from 04/01/2000 to 12/31/2021. These data come from the CDG Capital Bourse database. Stock liquidity is evaluated using three indicators to account for its multidimensional nature, including the displayed range (Qs), the Amihud illiquidity measure (Amh) and the zero return (Zr). The formulas for calculating these indicators are as follows:

$$Qs_{i,t} = \frac{(P_{i,t}^A - P_{i,t}^B)}{P_{i,t}^M} \quad (8)$$

$$Amh_{i,t} = \left( \sum_{t=1}^D \frac{\|r_{i,t}\|}{Vol_{i,t}} \right) / D_{i,t} \quad (9)$$

$$Zr_{i,t} = \begin{cases} 1, & r_{i,t} = 0 \\ 0, & r_{i,t} \neq 0 \end{cases} \quad (10)$$

With  $P_{i,t}^A$ ,  $P_{i,t}^B$ ,  $P_{i,t}^M$ ,  $r_{i,t}$  and  $Vol_{i,t}$  are the best ask price, best bid price, closing price, daily return, and daily volume of stock  $i$ , respectively. The displayed range  $Qs_{i,t}$  measures the depth of the market. The larger the spread between the best ask price and the best bid price, relative to the market price, the lower the liquidity of the stock [5]. Amihud's illiquidity ratio ( $Amh_{i,t}$ ) describes the change in daily price for a given trading volume; a low trading volume generating a higher return is synonymous with illiquidity. The zero return ( $Zr_{i,t}$ ) measures the number of days when the return is zero. It takes the value 1 if the return is zero and 0 otherwise. A high number of days of zero return is synonymous with low liquidity.

Each indicator is calculated daily per company; the aggregate daily indicator is an average of all companies. In sum, we have 5268 observations for Qs, 5469 observations for Amh and 5475 for Zr.

2) *Descriptive Statistics:* Fig. 4 shows that the measures of stock liquidity are chaotic, erratic, asymmetric, and non-linear to time. These characteristics are evidenced by the descriptive statistics presented in Table II. The liquidity measures are volatile as indicated by the coefficients of variation that are close to 0.5, and exceed 1 for the Amihud illiquidity ratio (Amh). Respectively, the Skewness and Kurtosis coefficients, indicate that the displayed range (Qs) and Amihud illiquidity ratio (Amh) are left skewed and pointed, while the zero return (Zr) is right skewed and flattened.

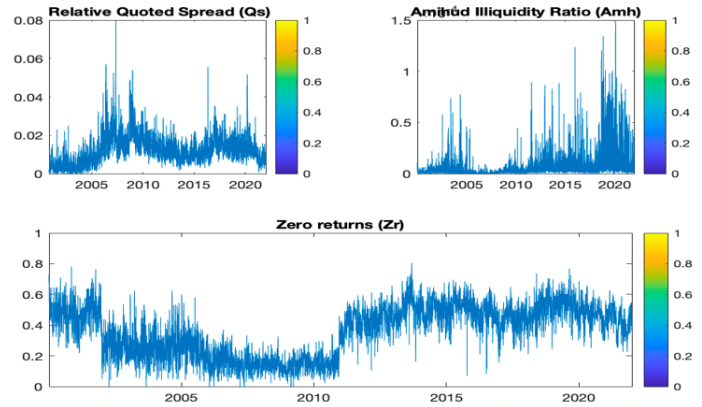


Fig. 4. Evolution of the Liquidity of Shares in the Casablanca Stock Exchange.

TABLE II. DESCRIPTIVE STATISTICS

	Qs	Amh	Zr
Number of observations	5268	5475	5475
Average	0,012	6,06E-06	0,362
Median	0,012	2,33E-06	0,397
Std	0,007	1,16E-05	0,167
Coefficient of variation	0,607	1,914	0,461
Kurtosis	6,512	34,004	1,874
Skewness	1,116	4,805	-0,196

#### B. Data Pre-Processing

The raw liquidity measures are first normalized by the Z-Score method; then denoised by the WT. However, since the choice of the decomposition method for the liquidity measures depends on the characteristics of the data, we first perform a time-frequency analysis of the liquidity measures. From Fig. 5, we can observe that Qs, Amh, and Zr experience low and medium frequency oscillations.

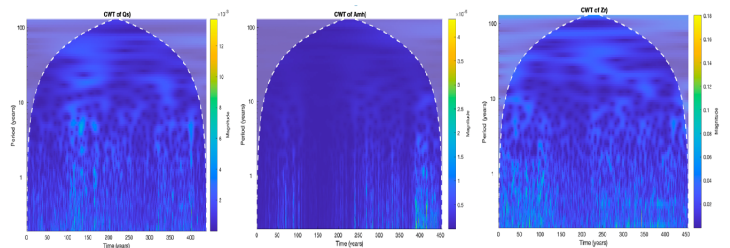


Fig. 5. Magnitudes Scalograms of Liquidity Measures.

Based on the time-frequency analysis while following the approach of [26], we present in III the methods used to decompose the stock liquidity measures.

TABLE III. MULTIREOLUTION ANALYSIS TECHNIQUES

Input signal	Signal characteristics					Decomposition technique
	Low frequency	Average frequency	Increase in frequency	Breaking	Trend	
Qs	Yes	Yes	No	No	No	VMD
Amh	Yes	Yes	No	No	No	VMD
Zr	Yes	Yes	No	Yes	No	MODWT

Since Qs and Amh experience low and medium frequency oscillations and peaks but no trend or break, the most appropriate method to decompose them is VMD. The latter decomposes the original signal into K intrinsic mode function (IMF) components. Fig. 6(a) shows the decomposition of Qs into five MFIs. The first IMFs (IMF1 to IMF3) locate low frequency oscillations, IMF4 locates mid-frequency oscillations and IMF 5 locates peaks. The Qs signal has experienced a few spikes between the year 2000 and 2021; the first ones are related to the effect of the 2007-2008 financial crisis while the last ones are due to COVID-19. Qs is noisy with midrange frequencies and spikes. Therefore, to reconstruct the noisy Qs signal, we neutralize IMF4 and IMF5. Fig. 6(a) highlights the original signal and the reconstructed signal.

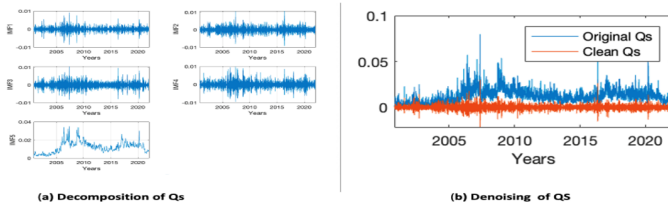


Fig. 6. Decomposition and Denoising of Qs.

Fig. 7(a) clarifies the decomposition of the Amh signal into five MFIs. This decomposition clearly shows that under the effect of COVID-19, the Amh signal experienced mid-frequency oscillations (MFI4) and spikes (MFI5) during 2020 and 2021. This has caused Amh to become very noisy. Thus, in order to denoise it, we reconstruct the signal while neglecting IMF1, IMF2, and IMF3. This reconstruction is shown in Fig. 7(b).

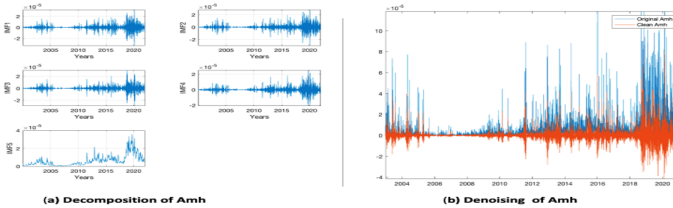


Fig. 7. Decomposition and Denoising of Amh.

Zr is decomposed by the MODWT method, as it experiences low and medium frequency oscillations and a jump in 2011. The MODWT method decomposes the original signal into wavelet coefficients and scaling coefficient. The wavelet coefficients identify high-frequency oscillations, while the scaling coefficients capture trends and jumps in a time series. Fig. 8(a) shows the MODWT algorithm’s decomposition of the Zr signal into five levels using the orthogonal Daubechies wavelet, with a level 1. We can observe from Fig. 8(a) that the Zr signal is noisy by medium frequency oscillations. Fig. 8(b) shows the reconstruction of the Zr signal after removing the noise.

1) Experimentation of the WT-CNN-GRU Model:

a) Presentation of the Empirical Results: In subsection III-B1, Qs and Amh were decomposed by the VMD method while Zr was decomposed by the MODWT method.

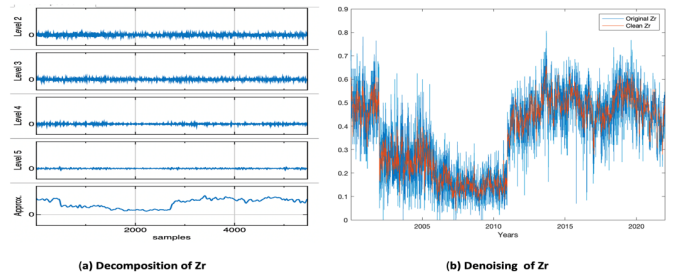


Fig. 8. Decomposition and Denoising of Zr.

After the decomposition and denoising operations, we have 5469 MFIs for Amh, 5268 MFIs for Qs and 5475 approximation coefficients for Zr. The selected MFIs and approximation coefficients constitute inputs for the CNN-GRU model. The data for these inputs are further splitted into 80% training data and 20% test data. Table IV details this split.

TABLE IV. TRAINING AND TEST DATA

Liquidity measurement	Training data		Test data	
	Number of observations	Period	Number of observations	Period
Qs	4215	November 01, 2000 to October 03, 2017	1053	October 04, 2017 to December 31, 2021
Amh	4370	January 04, 2000 to July 19, 2017	1099	July 20, 2017 to December 31, 2021
Zr	4380	January 04, 2000 to July 31, 2017	1095	August 01, 2017 to December 31, 2021

The CNN-GRU model is trained according to the parameters described in Table I. A comparative study between the forecasted and actual values of the three denoised liquidity measures is shown in Fig. 9 to Fig. 11. We can observe that the forecasted values are almost equal to the actual values and that our model is able to make accurate forecasting even under sudden shocks, such as the case of COVID-19 in 2020 and 2021. From Fig. 9 to Fig. 11, we can also notice that the number of outliers in the test errors of the proposed model is very small.

Table V lists the metrics for evaluating the predictive performance of test data. From Table V, we can notice that the proposed model gives very low MAE, MAPE, and RMSE, Theil’s U statistics less than 1, and ( $R^2$ ) close to 1. Our model demonstrates excellent forecasting performance compared to models proposed in previous studies.

The author in [5] proposed a NARX neural network to predict the liquidity of stocks listed on the Casablanca Stock Exchange. Their model is evaluated based on MSE which indicates a value of 0.0083 for Qs and 0.0023 for Zr. The author in [6] estimates that the LSTM model is more efficient than the linear regression method and the multilayer perceptron. The results of this study indicate that the LSTM model exhibits better MSEs. The MSE of the Amihud ratio indicates a value of 0.0169 and 0.0252 on the Ho Chi Min and Hanoi Stock Exchange respectively.

Statistically, our model far exceeds the performance of models postulated by previous studies. Moreover, these models are only tested on a small number of observations. The authors in [5] and [6] only tested their model on the basis of 12 and 110 observations respectively, whereas our model is tested on 1053 observations for Qs, 1099 for Amh, and 1095 for Zr. As a result, the MSEs of earlier models are less reliable. The commonality of previous studies is the use of MSE as an endpoint. However, the latter is more sensitive to outliers. On the contrary, our model is evaluated on the basis of multiple indicators.

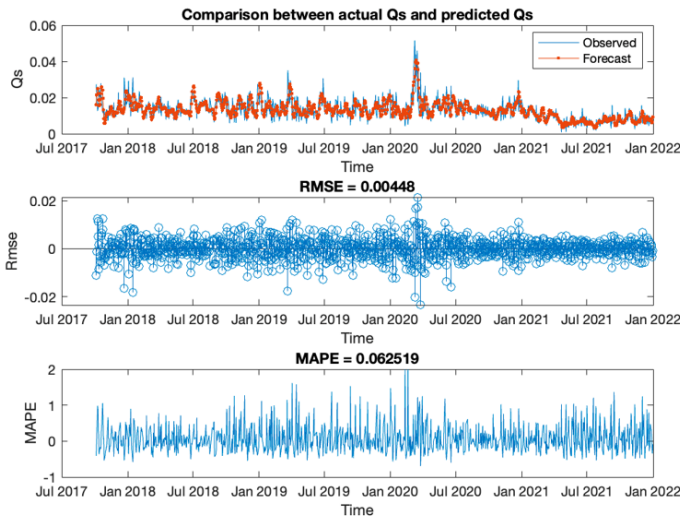


Fig. 9. Comparison between Actual Qs and Qs Predicted by the WT-CNN-GRU Model.

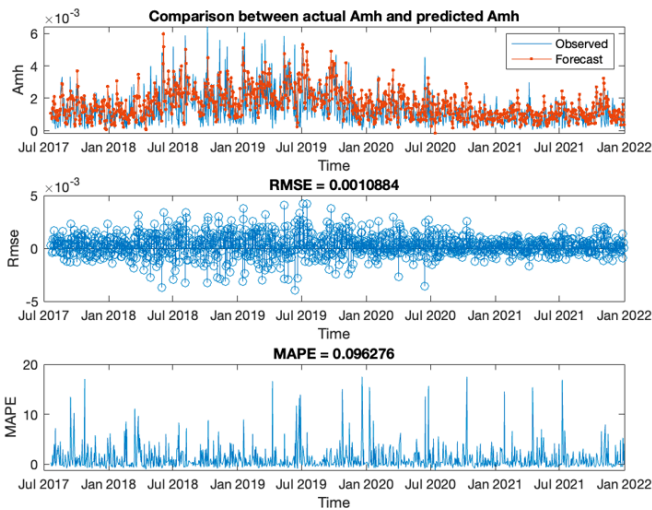


Fig. 10. Comparison between Actual Amh and Amh Predicted by the WT-CNN-GRU Model.

*b) Comparison with Similar and Alternative Models:*

To prove the effectiveness of the proposed model, we compare its performance to similar models, such as WT-CNN-LSTM and WT-CNN-BILSTM. In addition, since the focus of our study is whether the introduction of WT improves the predictive ability of deep neuron networks, the proposed

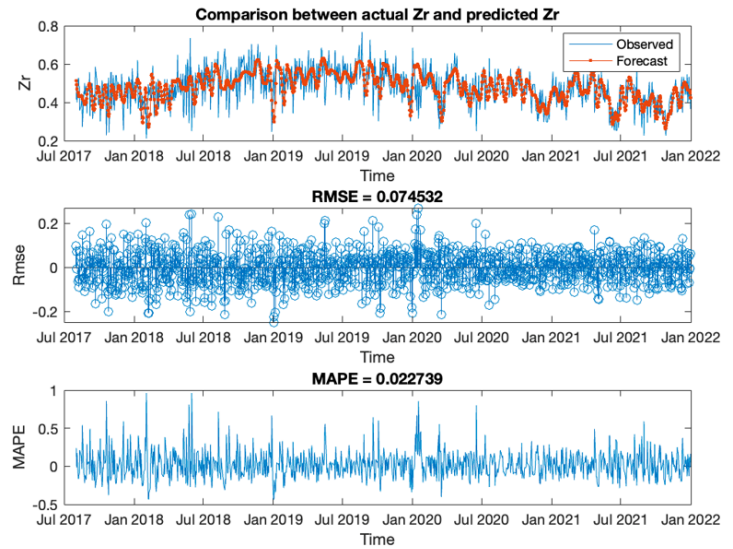


Fig. 11. Comparison between Actual Zr and Zr Predicted by the WT-CNN-GRU Model.

TABLE V. INDICATORS FOR EVALUATING THE PREDICTIVE PERFORMANCE OF THE CNN-GRU MODEL

Liquidity measurement	MAE	MAPE	RMSE	Theil's U	R <sup>2</sup>
Qs	3.1e-05	0.062	0.004	0.072	0.7379
Amh	0.0002	0.096	0.001	0.037	0.6325
Zr	-0.0025	0.022	0.074	0.079	0.7880

model is compared to alternative models such as CNN-GRU, CNN-LSTM and CNN-BILSTM. While the inputs of the similar models (WT-CNN-LSTM and WT-CNN-BILSTM) are denoised data, those of the alternative models (CNN-LSTM, CNN, BILSTM, and CNN-GRU,) are normalized raw data.

As showcased in Table VI, if we compare the proposed WT-CNN-GRU model with the CNN-GRU model, we can notice that the denoising of the data by WT improved the forecasting results of stock liquidity. At the level of Qs, the WT was able to reduce MAPE and Theil's U by 33% and 57% respectively, and increase R<sup>2</sup> by 48%. In addition, denoising the Amh data contributed to the improvement of MAE, MAPE, RMSE, Theil's U and R<sup>2</sup> by 154%, 84%, 21%, 94% and 40% respectively. Finally, the denoising of Zr data by WT strongly improved MAE, MAPE, RMSE, Theil's U and R<sup>2</sup> by 14%, 72%, 24%, 9% and 97% respectively. These improved results are obtained not only from the denoising of the data, but also from the appropriate choice of the decomposition methods.

In Table VI, we can also notice that the proposed model WT-CNN-GRU also outperforms similar models, such as WT-CNN-LSTM and WT-CNN-BILSTM. The GRU closed cell is therefore the most effective hybrid model when it is compared to the LSTM and BILSTM cells. GRU contains several hidden layers, which can efficiently identify the fluctuation characteristics of liquidity measures. GRU also optimizes the network structure and reduces information redundancy.



TABLE VI. EVALUATION CRITERIA VALUES FOR DIFFERENT FORECASTING MODELS

	Models	MAE	MAPE	RMSE	Theil's U	R <sup>2</sup>
Qs	WT-CNN-LSTM	6.71e-05	0.09	0.008	0.1619	0.7174
	WT-CNN-BILSTM	0.0001	0.117	0.004	0.1834	0.4953
	WT-CNN-GRU	3.1e-05	0.062	0.004	0.0723	0.7379
	CNN-LSTM	0.0004	0.07	0.005	0.2143	0.1063
	CNN-BILSTM	8.6e-04	0.095	0.005	0.2056	0.0674
	CNN-GRU	2.3e-05	0.0933	0.004	0.1683	0.4961
Amh	WT-CNN-LSTM	0.0024	0.1479	0.0011	0.3694	0.3639
	WT-CNN-BILSTM	2.7e-05	0.3631	0.0011	0.3840	0.0223
	WT-CNN-GRU	0.0002	0.096	0.0011	0.0371	0.6325
	CNN-LSTM	0.0001	0.2558	0.0015	0.5295	0.0092
	CNN-BILSTM	0.0001	0.2937	0.0010	0.3704	0.0842
	CNN-GRU	-3.7e-04	0.5907	0.0014	0.6524	-0.4526
Zr	WT-CNN-LSTM	-0.0025	0.0226	0.07082	0.0741	0.6104
	WT-CNN-BILSTM	0.0040	0.0190	0.07482	0.0778	0.33970
	WT-CNN-GRU	-0.0025	0.022	0.074	0.079	0.7880
	CNN-LSTM	-0.028	0.0851	0.0849	0.0913	0.2285
	CNN-BILSTM	-0.0074	0.0473	0.0861	0.0906	0.1020
	CNN-GRU	0.0022	0.0793	0.0976	0.0865	0.3992

#### IV. CONCLUSION

Forecasting equity liquidity is crucial for investors, issuers, and financial market regulators. As a financial series, stock liquidity is non-stationary, non-linear, chaotic, and noisy. Therefore, it is very difficult to accurately forecast inventory liquidity. The purpose of this study is to propose a model capable of effectively predicting inventory liquidity. Inspired by the hybrid research stream in financial time series forecasting, we proposed a WT-CNN-GRU model.

By testing it on all stocks listed on the B.V.C, our model showed excellent forecasting performance compared to models from previous studies and other similar or alternative models, such as WT-CNN-LSTM, WT-CNN -BILSTM, CNN-GRU, CNN-LSTM, and CNN-BILSTM. These improved performances are jointly explained by the neural networks WT, CNN, and the closed cell GRU. Choosing the right method for data decomposition and denoising was key to improving the results. The CNN was used to capture features that the WT did not capture, and they worked together to denoise the data. The GRU cell captured the time dependencies of stock liquidity, which has an advantage over the LSTM or BILSTM cell due to its ability to quickly catch time dependencies while avoiding information redundancy.

This study has three contributions. First, it is considered to be one of the few studies that have addressed the issues of forecasting inventory liquidity. Second, unlike previous studies that used a single data denoising method, we opted for methods consistent with our data. Third, the proposed model can predict stock liquidity even in the face of adverse shocks, such as the COVID-19 pandemic.

This study contributes to the enrichment of the forecast field of the financial series. It can be a useful analytical framework capable of helping investors, issuers, and financial market regulators predict stock liquidity. The model proposed in this study is also extended to predict other financial risks.

However, to verify its robustness, we suggest that future researchers test it in other emerging and developed markets. We also believe that our results can be improved by integrating into the proposed model global and specific exogenous variables for companies listed on stock exchanges.

#### REFERENCES

- [1] T. Chordia, L. Shivakumar, et A. Subrahmanyam, "Liquidity dynamics across small and large firms", *Econ. Notes*, vol. 33, no 1, p. 111-143, 2004.
- [2] R. Ma, H. D. Anderson, et B. R. Marshall, "Market volatility, liquidity shocks, and stock returns: Worldwide evidence", *Pac.-Basin Finance J.*, vol. 49, p. 164-199, 2018.
- [3] N. Cakici et A. Zaremba, "Liquidity and the cross-section of international stock returns", *J. Bank. Finance*, vol. 127, p. 106123, 2021.
- [4] T. Zhang et S. H. Lence, "Liquidity and asset pricing: Evidence from the Chinese stock markets", *North Am. J. Econ. Finance*, vol. 59, p. 101557, 2022.
- [5] A. K. Mishra, B. Parikh, et R. W. Spahr, "Stock market liquidity, funding liquidity, financial crises and quantitative easing", *Int. Rev. Econ. Finance*, vol. 70, p. 456-478, 2020.
- [6] A. K. Mishra, B. Parikh, et R. W. Spahr, "Contemporaneous linkages: Funding liquidity and stock market spirals", *Int. J. Finance Econ.*, vol. 26, no 4, p. 5912-5929, 2021.
- [7] T. Geithner, "Liquidity and financial markets", 2007.
- [8] C. G. Rösch et C. Kaserer, "Reprint of: Market liquidity in the financial crisis: The role of liquidity commonality and flight-to-quality", *J. Bank. Finance*, vol. 45, p. 152-170, 2014.
- [9] J. A. Skjeltorp et B. A. Ødegaard, "When do listed firms pay for market making in their own stock?", *Financ. Manag.*, vol. 44, no 2, p. 241-266, 2015.
- [10] U. Shahzad, J. Liu, et F. Luo, "Stock liquidity and corporate trade credit strategies: evidence from China", *J. Bus. Econ. Manag.*, vol. 23, no 1, p. 40-59-40-59, 2022.
- [11] S. Stereńczak et J. Kubiak, "Dividend policy and stock liquidity: Lessons from Central and Eastern Europe", *Res. Int. Bus. Finance*, vol. 62, p. 101727, 2022.
- [12] J. Brogaard, D. Li, et Y. Xia, "Stock liquidity and default risk", *J. Financ. Econ.*, vol. 124, no 3, p. 486-502, 2017.
- [13] H. H. Trinh, C. P. Nguyen, W. Hao, et U. Wongchoti, "Does stock liquidity affect bankruptcy risk? DID analysis from Vietnam", *Pac.-Basin Finance J.*, vol. 69, p. 101634, 2021.
- [14] Y. Amihud, A. Hameed, W. Kang, et H. Zhang, "Stock liquidity and the cost of equity capital in global markets", *J. Appl. Corp. Finance*, vol. 27, no 4, p. 68-74, 2015.
- [15] M. B. Houad et Y. Oubouali, "Forecasts of the liquidity of shares listed on the Casablanca Stock Exchange. Comparison between ARIMA modeling and NARX neural networks" "Prévisions de la liquidité des actions cotées à la bourse des valeurs de Casablanca. Comparaison entre la modélisation ARIMA et les réseaux de neurones NARX", *Rev. Gest. Organ.*, vol. 10, no 2, p. 83-99, 2018.
- [16] P. Q. Khang et al., "Machine learning for liquidity prediction on Vietnamese stock market", *Procedia Comput. Sci.*, vol. 192, p. 3590-3597, 2021.
- [17] M. Vogl, P. G. Rötzel, et S. Homes, "Forecasting performance of wavelet neural networks and other neural network topologies: A comparative study based on financial market data sets", *Mach. Learn. Appl.*, vol. 8, p. 100302, 2022.
- [18] H. Malik, A. K. Yadav, F. P. G. Márquez, et J. M. Pinar-Pérez, "Novel application of Relief Algorithm in cascaded artificial neural network to predict wind speed for wind power resource assessment in India", *Energy Strategy Rev.*, vol. 41, p. 100864, 2022.
- [19] F. Rodríguez, I. Azcárate, J. Vadiello, et A. Galarza, "Forecasting intra-hour solar photovoltaic energy by assembling wavelet based time-frequency analysis with deep learning neural networks", *Int. J. Electr. Power Energy Syst.*, vol. 137, p. 107777, 2022.
- [20] C. Song, L. Yao, C. Hua, et Q. Ni, "A novel hybrid model for water quality prediction based on synchro-squeezed wavelet transform technique and improved long short-term memory", *J. Hydrol.*, vol. 603, p. 126879, 2021.
- [21] Q. Gu, Y. Chang, N. Xiong, et L. Chen, "Forecasting Nickel futures price based on the empirical wavelet transform and gradient boosting decision trees", *Appl. Soft Comput.*, vol. 109, p. 107472, 2021.

- [22] M. Risse, "Combining wavelet decomposition with machine learning to forecast gold returns", *Int. J. Forecast.*, vol. 35, no 2, p. 601-615, 2019.
- [23] X. Liu, H. Liu, Q. Guo, et C. Zhang, "Adaptive wavelet transform model for time series data prediction", *Soft Comput.*, vol. 24, no 8, p. 5877-5884, 2020.
- [24] S. C. Olhede et A. T. Walden, "Generalized morse wavelets", *IEEE Trans. Signal Process.*, vol. 50, no 11, p. 2661-2670, 2002.
- [25] J. Chung, C. Gulcehre, K. Cho, et Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling", *ArXiv Prepr. ArXiv14123555*, 2014.
- [26] J. Gilles, "Empirical wavelet transform", *IEEE Trans. Signal Process.*, vol. 61, no 16, p. 3999-4010, 2013.

# Visual Navigation System for Autonomous Drone using Fiducial Marker Detection

Mohammad Soleimani Amiri

Department of Mechanical and Manufacturing Engineering,  
Faculty of Engineering and Built Environment,  
Universiti Kebangsaan Malaysia,  
Bangi 43600, Selangor, Malaysia

Rizauddin Ramli

Department of Mechanical and Manufacturing Engineering,  
Faculty of Engineering and Built Environment,  
Universiti Kebangsaan Malaysia,  
Bangi 43600, Selangor, Malaysia

**Abstract**—Drones have been quickly developing for civilian applications in recent years. Because of the nonlinearity of the mathematical drone model, and the importance of precise navigation to avoid possible dangers, it is necessary to establish an algorithm to localize the drone simultaneously and maneuver it to the desired destination. This paper presents a visual-based multi-stage error tolerance navigation algorithm of an autonomous drone by a tag-based fiducial marker detection in finding its target. Dynamic and kinematic models of the drone were developed by Newton-Euler. The position and orientation of the drone, related to the tag, are determined by AprilTag, which is used as feedback in a closed-loop control system with an Adjustable Proportional-Integral-Derivative (APID) controller. Parameters of the controller are tuned based on steady-State error, which is defined as the distance of the drone from the desired point. The sequence of path trajectory, that drone follows to reach the desired point, is defined as a navigation algorithm. A model of the drone was simulated in a virtual outdoor to mimic hovering in complex obstacles environment. The results present satisfactory performance of the navigation system programmed by the APID controller in comparison with the conventional Proportional-Integral-Derivative (PID) controller. It can be ascertained that the proposed navigation system based on a tag marker in the closed-loop control system is applicable to maneuvering the drone autonomously and useful for various industrial tasks in indoor/outdoor environments.

**Keywords**—Proportional-Integral-Derivative (PID) controller; AprilTag detection system; autonomous navigation; fiducial marker detection

## I. INTRODUCTION

The drone has the advantage of reaching high locations and extreme environments that are difficult to be accessed by humans and other ground vehicles [1], [2], [3]. Since its deployment to industrial applications, drones have attracted a lot of attention in various industries [4], [5], [6]. For instance, Huang et al. [7] presented for scene detection by high-resolution imagery adopted through autonomous drone navigation using landmark detection and recognition

The drone can be navigated autonomously or by a human pilot. There are several works about developing the navigation algorithm for the drone [8], [9], [10], [11]. Hodge et al. [12] presented a generic navigation algorithm that utilizes onboard sensors' data of the drone to navigate the drone to the target. Miranda et al. [13] developed a navigation system for autonomous drones that generates a path between a start and a final point and controls the drone to follow this path.

Tang et al. [14] studied an algorithm based on a multi-sensor system including multiple cameras and a 2-D laser scanner using the AprilTag target for drone application. Malyuta et al. [15] presented an autonomous drone for precision agriculture applications by employing the drone, flying through the farmland, for long-term monitoring missions without any human supervision by using AprilTag for localization. Lee et al. [16] established drone navigation by mapping the definitions of vehicular autonomy levels to specific drone tasks in order to create a clear definition of autonomy when applied to drones.

From literature [12], [13], [14], [15], [16], there are various types of strategies to increase the performance of navigation of the drone. Developing a simpler navigation strategy with more efficient performance has attracted much attention from researchers. The contributions of this paper are given as follows:

- The kinematics and dynamics of the drone are determined.
- The visual-based navigation and Adjustable Proportional-Integral-Derivative (APID), in which the actual position of the drone is captured by Fiducial Marker Detection, are investigated.

This paper is organized as follows: dynamic models of a drone are expressed in Section II. Section III addresses the development of APID and visual-based navigation. Section IV represents the performance and result of the proposed APID and navigation strategy in real-time navigation of a drone model in the virtual environment interacting with Robot Operating System (ROS). The conclusion is mentioned in Section V.

## II. DYNAMICS OF DRONE

The mathematical model of the drone is presented as follows [17]:

$$\dot{\eta} = \eta\nu \quad (1)$$

where the vector  $\eta$  represents position and orientation in the earth or inertial frame shown in Fig. 1 as follows:



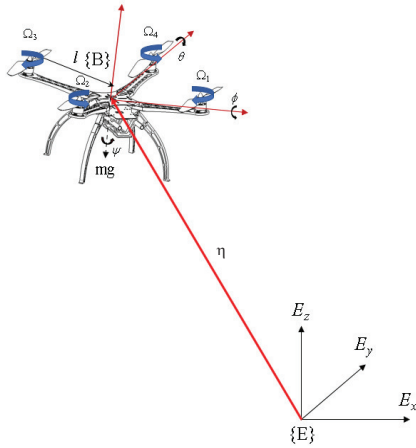


Fig. 1. Drone Model Coordinate Systems.

$$\eta = |x \ y \ z \ \phi \ \theta \ \psi|^T \quad (2)$$

And the equation of motion for the drone can be represents as,

$$M_{RB}\dot{\nu} + C_{RB}(\nu)\nu = \tau \quad (3)$$

where  $M_{RB}$  is the rigid-body inertia matrix,  $C_{RB}(\nu)$  represents the centripetal and Coriolis terms and  $\tau$  is the generalized forces and torques, all of which are expressed in the body-fixed frame.

And vector  $\nu$  represents the linear velocities and angular velocities of the drone in the body-fixed frame as follows:

$$\nu = \begin{vmatrix} V^B \\ \omega^B \end{vmatrix} = |u \ v \ w \ p \ q \ r|^T \quad (4)$$

where  $V^B$  is the linear velocities in the body ( $B$ ) frame and  $\omega^B$  is the angular velocities in the body frame respectively. In 3 dimension rigid body dynamics, the two reference frames are linked by the linear and angular velocities as,

$$\nu = R(\eta)V^B \quad (5)$$

$$\omega = T(\eta)\omega^B \quad (6)$$

In general,  $j(\eta)$  is the transformation matrix of the orientation and position of the body frame with respect to the inertial frame of the drone can be shown as [18],

$$j(\eta) = \begin{vmatrix} R(\eta) & 0_{3 \times 3} \\ 0_{3 \times 3} & T(\eta) \end{vmatrix} \quad (7)$$

The transformation matrix that indicates the relation between angular velocities in the body frame and angular velocities in the inertial frame is represented as follows,

$$T(\eta) = \begin{vmatrix} 1 & s_\phi t_\theta & c_\phi t_\theta \\ 0 & c_\phi & -s_\phi \\ 0 & s_\phi/c_\theta & c_\phi/c_\theta \end{vmatrix} \quad (8)$$

The transformation matrix  $j(\eta)$  with the rotation matrix  $R(\eta)$  which describes the relationship between the linear velocities in the body-fixed frame and the linear velocities in the inertial frame is shown as,

$$R(\eta) = R_B^E(\psi)R_B^E(\theta)R_B^E(\phi) \quad (9)$$

$$R(\eta) = \begin{vmatrix} c_\psi c_\theta & -s_\psi c_\theta + c_\psi s_\theta s_\phi & s_\psi s_\theta + c_\psi c_\theta s_\phi \\ s_\psi c_\theta & c_\psi c_\theta + s_\psi s_\theta s_\phi & -c_\psi s_\theta + s_\psi c_\theta s_\phi \\ -s_\theta & c_\theta s_\phi & c_\theta c_\phi \end{vmatrix} \quad (10)$$

Since the vector of linear and angular velocities  $\dot{\eta}$  in earth frame can be expressed as,

$$\dot{\eta} = \begin{vmatrix} \dot{x} \\ \dot{y} \\ \dot{z} \\ \dot{\phi} \\ \dot{\theta} \\ \dot{\psi} \end{vmatrix} \quad (11)$$

The kinematics relation between the linear and angular velocities of the body-fixed frame related to the inertial frame is described by

$$\dot{\eta} = \begin{vmatrix} u(s_\psi c_\theta) + v(s_\psi s_\theta s_\phi + c_\psi c_\theta) + w(c_\psi s_\theta c_\phi + s_\psi s_\phi) \\ u(c_\psi c_\theta) + v(c_\psi s_\theta s_\phi - s_\psi c_\theta) + w(s_\psi s_\theta c_\phi - c_\psi s_\phi) \\ u(-s_\theta) + v(c_\theta s_\phi) + w(c_\theta c_\phi) \\ p + q(s_\psi t_\theta) + r(c_\psi t_\theta) \\ q(c_\psi) - r(s_\psi) \\ q\left(\frac{s_\psi}{c_\theta}\right) + r\left(\frac{c_\psi}{c_\theta}\right) \end{vmatrix} \quad (12)$$

By assuming that the center of mass is fixed at the center of the origin of the body frame and the body has rotational symmetry around the center of mass, the resulting rigid-body inertia matrix  $M_{RB}$  is presented as,

$$M_{RB} = \begin{vmatrix} mI_{3 \times 3} & 0_{3 \times 3} \\ 0_{3 \times 3} & I \end{vmatrix} = \begin{vmatrix} m & 0 & 0 & 0 & 0 & 0 \\ 0 & m & 0 & 0 & 0 & 0 \\ 0 & 0 & m & 0 & 0 & 0 \\ 0 & 0 & 0 & I_{xx} & 0 & 0 \\ 0 & 0 & 0 & 0 & I_{yy} & 0 \\ 0 & 0 & 0 & 0 & 0 & I_{zz} \end{vmatrix} \quad (13)$$

where  $m$  is the total mass of the drone and  $I_{xx}$ ,  $I_{yy}$  and  $I_{zz}$  are the moments of inertia [19]. By using Newton-Euler equation, the drone rigid body are affected by external forces  $F^B$  and torques  $\tau^B$  as,

$$\begin{vmatrix} mI_{3 \times 3} & 0_{3 \times 3} \\ 0_{3 \times 3} & I \end{vmatrix} \begin{vmatrix} \dot{V}^B \\ \dot{\omega}^B \end{vmatrix} + \begin{vmatrix} \omega^B \times mV^B \\ \omega^B \times I\omega^B \end{vmatrix} = \begin{vmatrix} F^B \\ \tau^B \end{vmatrix} \quad (14)$$

Expanding the equation above, we obtain,

$$\begin{pmatrix} m & 0 & 0 & 0 & 0 & 0 \\ 0 & m & 0 & 0 & 0 & 0 \\ 0 & 0 & m & 0 & 0 & 0 \\ 0 & 0 & 0 & I_{xx} & 0 & 0 \\ 0 & 0 & 0 & 0 & I_{yy} & 0 \\ 0 & 0 & 0 & 0 & 0 & I_{zz} \end{pmatrix} \begin{pmatrix} \dot{u} \\ \dot{v} \\ \dot{w} \\ \dot{p} \\ \dot{q} \\ \dot{r} \end{pmatrix} + \begin{pmatrix} -rmv + qmw \\ rmu - pmw \\ -qmu + pmv \\ -rI_{yy}q + qI_{zz}r \\ -rI_{xx}p + pI_{zz}r \\ qI_{xx}p - pI_{yy}q \end{pmatrix} = \begin{pmatrix} F_x^B \\ F_y^B \\ F_z^B \\ \tau_x^B \\ \tau_y^B \\ \tau_z^B \end{pmatrix} \quad (15)$$

The external forces and torques for each component of x, y and z axis can be determined as,

$$F_x^B = m(\dot{u} - rv + qw) \quad (16)$$

$$F_y^B = m(\dot{v} + ru - pw) \quad (17)$$

$$F_z^B = m(\dot{w} - qu + pv) \quad (18)$$

$$\tau_x^B = I_{xx}\dot{p} - rq(I_{yy} - I_{zz}) \quad (19)$$

$$\tau_y^B = I_{yy}\dot{q} + rp(I_{xx} - I_{zz}) \quad (20)$$

$$\tau_z^B = I_{zz}\dot{r} - qp(I_{xx} - I_{yy}) \quad (21)$$

The centripetal and Coriolis terms are represented by the matrix

$$C_{RB}(\nu)\nu = \begin{pmatrix} 0 & 0 & 0 & 0 & mw & -mv \\ 0 & 0 & 0 & -mw & 0 & mu \\ 0 & 0 & 0 & mv & -mu & 0 \\ 0 & 0 & 0 & 0 & I_{zz}r & -I_{yy}q \\ 0 & 0 & 0 & -I_{zz} & 0 & I_{xx}p \\ 0 & 0 & 0 & I_{yy}q & -I_{xx}p & 0 \end{pmatrix} \begin{pmatrix} u \\ v \\ w \\ p \\ q \\ r \end{pmatrix} \quad (22)$$

The generalized forces  $\tau$  can be divided into three components,

$$\tau = \tau_{gravitational} + \tau_{damping} + \tau_{actuators} = G(\eta) + D(\nu) + \tau_c(u) \quad (23)$$

where  $G(\eta)$  is the gravitational component,  $D(\nu)$  is the damping component,  $\tau_c(u)$  is the forces generated by the actuators and  $u$  is the input vector to the drone's motors.

The gravitational component points downwards along the z-axis with respect to the earth frame which in the body frame corresponds to:

$$G(\eta) = \begin{pmatrix} R_E^B & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 0 \\ 0 \\ -mg \\ 0 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} -mgs_\theta \\ mgc_\theta \sin\phi \\ mgc_\theta c_\phi \\ 0 \\ 0 \\ 0 \end{pmatrix} \quad (24)$$

where  $R_E^B$  denotes the inverse of  $R(\eta)$  as rotation matrix of the body-fixed frame relative to inertial frame.

$$R_E^B = \begin{pmatrix} c_\psi c_\theta & s_\psi c_\theta & -s_\theta \\ c_\psi s_\theta s_\phi - s_\psi c_\phi & s_\psi s_\theta s_\phi + c_\psi c_\theta & c_\theta s_\phi \\ c_\psi s_\theta c_\phi + s_\psi s_\phi & s_\psi s_\theta c_\phi - c_\psi c_\phi & c_\theta c_\phi \end{pmatrix} \quad (25)$$

The damping component is the linear matrix

$$D(\nu) = D_0\nu = \begin{pmatrix} D_{0,u}u \\ D_{0,v}v \\ D_{0,w}w \\ D_{0,p}p \\ D_{0,q}q \\ D_{0,r}r \end{pmatrix} \quad (26)$$

The forces and torques generated by the actuators are assumed to be linear in the input,

$$\tau_c(u) = u = \begin{pmatrix} 0 \\ 0 \\ u_{trottle} \\ u_{roll} \\ u_{pitch} \\ u_{yaw} \end{pmatrix} \quad (27)$$

$u_{trottle}$  is the control input of the trust force as the force affecting the velocity  $w$  in the z-direction. The force of trust gives a lifting power that makes the drone flies and depends on the sum of the speed of the four propellers. Since the rotors are fixed their total thrust will always pull upwards along the z-axis of the body frame.

$$u_{trottle} = k_f \sum_{i=1}^4 \Omega_i^2 = k_f(\Omega_1^2 + \Omega_2^2 + \Omega_3^2 + \Omega_4^2) \quad (28)$$

where  $i$  is the number of motors (propeller),  $\Omega_i$  is the speed of motor  $i$  and,  $k_f$  is thrust constant. Next, by increasing or decreasing the speed of the four rotors independently, it will create torques around the x-y-z axes and thus create roll-,pitch- and yaw-rotations. By always decreasing the speed of one rotor as much as increasing the speed of another the total thrust is retained. Therefore, the control input of roll  $u_{roll}$  and  $u_{pitch}$  are related to the torque and can be obtained by multiplying the force by the distance and the rotors will affect the total rotation about a certain axis differently depending on the distance from the center of gravity(CoG) of the drone.

$$u_{roll} = k_f l(-\Omega_2^2 + \Omega_4^2) \quad (29)$$

$$u_{pitch} = k_f l(\Omega_1^2 - \Omega_3^2) \quad (30)$$

$$u_{yaw} = k_M(\Omega_1^2 - \Omega_2^2 + \Omega_3^2 - \Omega_4^2) \quad (31)$$

where  $k_f$  is the coefficient of the force affecting the velocity  $w$  in the  $z$ -direction related to the thrust constant of the drone. Meanwhile,  $l$  is the distance between the axis of rotation of each rotor to the origin of the body reference frame which should coincide with the CoG of the drone and  $k_M$  is the moment constants, respectively.

### III. DEVELOPMENT OF APID FOR VISUAL BASED NAVIGATION

In order to maneuver the drone autonomously, based on the position and orientation of the tag marker to the drone body frame, a closed-loop control system with APID is occupied, in which its parameters are tuned by an Adjustment Mechanism (AM) [20], [21]. PID has been regarded as one of the most popular controllers in the industry because of its ease of implementation, and efficient performance [22], [23], [24]. Fig. 2 shows the control system of AprilTag navigation.

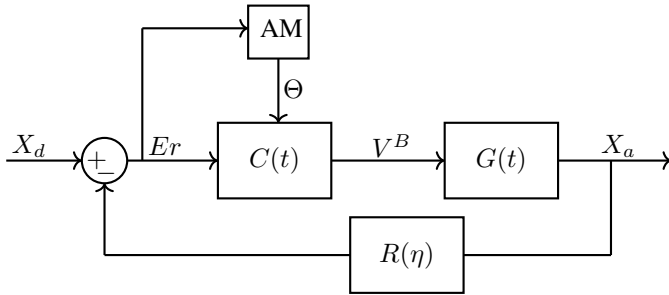


Fig. 2. Block Diagram of the Control System.

The output of AM,  $\Theta = [K_p \ K_i \ K_d]$ , is a matrix of APID parameters. The steady-State error is determined as the difference between the actual and desired position of the drone, given as follows:

$$E_r = X_d - X_a \quad (32)$$

where  $E_r = [e_{r_x} \ e_{r_y} \ e_{r_z} \ e_{r_\phi}]^T$

is the error matrix;  $X_d = [x_d \ y_d \ z_d \ \phi_d]^T$

is desired matrix;  $X_a = [x_a \ y_a \ z_a \ \phi_a]^T$

is the actual path matrix. The output of the APID is the velocity of the body frame, which is influenced by the proportion, integral, and derivative of the steady-State error. The mathematical model of the controller is expressed as follows:

$$C(t) = K_p E_r + K_i \int E_r dt + K_d \frac{dE_r}{dt} \quad (33)$$

$G(t)$  is the equation of motion for the body frame of the drone regarding the earth frame as a rigid body. The actual data is captured by the camera, which is attached to the body of the drone. Because the coordinate of the AprilTag marker  $\{t_i\}$  and the drone's body frame  $\{B\}$  does not match, a rotation

matrix is needed to align the axes of the tag marker with the axes of drone movement [25], as follows:

$$R_t^B = -R_t^B(\theta)R_t^B(\phi) \quad (34)$$

where  $R_t^B(\theta) \in R^3$  and  $R_t^B(\phi) \in R^3$  are pitch and roll rotation matrix of tag frame  $\{t\}$  to drone body frame  $\{B\}$ . Fig. 3 represents schematic of drone navigation based on tag marker.

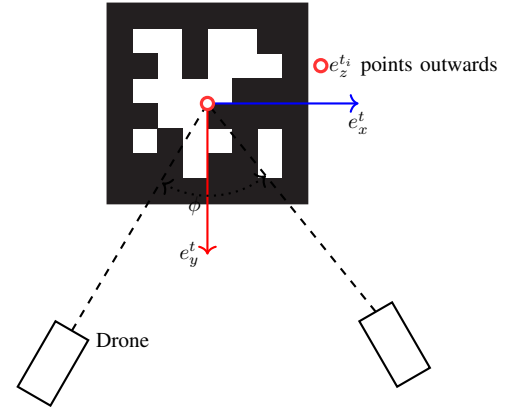


Fig. 3. Schematic of Drone Navigation based on Tag Marker.

The parameters of APID controller are tuned in such a way to prevent overshoot to avoid drone collision and achieve it slowly. Therefore AM is introduced for tuning of APID controller, in which three rules for distances and angle have been determined. Three sets of APID parameters have been established as follows:

$$\Theta_{s1} = [0.05 \ 0.01 \ 0.009] \quad (35)$$

$$\Theta_{s2} = [0.1 \ 0.05 \ 0.01] \quad (36)$$

$$\Theta_{s3} = [0.2 \ 0.1 \ 0.02] \quad (37)$$

Table I represents the rules for AM to tune APID controllers and sets of APID parameters and related error range.

TABLE I. RULES OF AM FOR APID PARAMETERS AND ERROR RANGES

Error range	Category	PID parameters
$0(m) \leq e_{r_{x,y,z}} < 0.2(m)$	N	$\Theta_{s1}$
$0.2(m) \leq e_{r_{x,y,z}} < 1(m)$	AN	$\Theta_{s2}$
$1(m) \leq e_{r_{x,y,z}} < D_{max}(m)$	F	$\Theta_{s3}$
$ e_{r_\phi}  \leq 20^\circ$	SA	$\Theta_{s1}$
$ e_{r_\phi}  \geq 20^\circ \ \& \  e_{r_\phi}  \leq 45^\circ$	MA	$\Theta_{s2}$
$ e_{r_\phi}  \geq 20^\circ \ \& \  e_{r_\phi}  \geq 45^\circ \ \& \  e_{r_\phi}  \leq A_{max}^\circ$	WA	$\Theta_{s3}$

$D_{max}(m)$  and  $A_{max}^\circ$  are the maximum distance and angle that can be recognized by AprilTag system. Fig. 4 represents a flowchart of rules for AM to adjust the APID parameters.

The scenario of the navigation algorithm is divided into four stages. Firstly the control system is applied to uplift

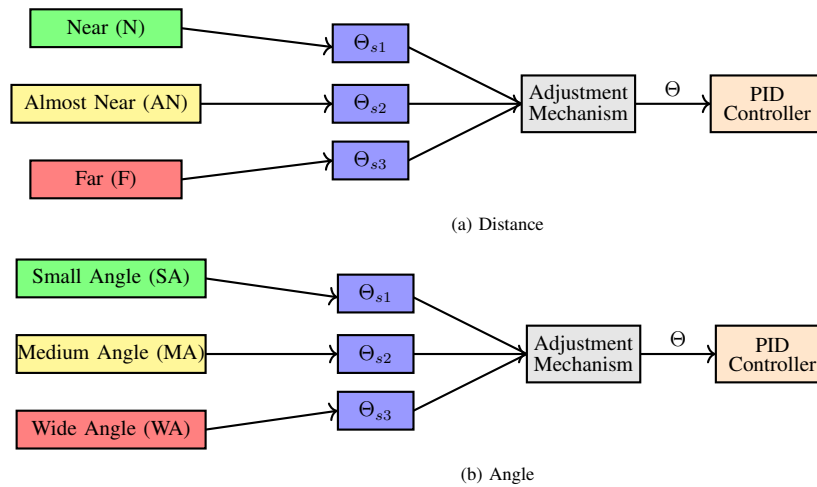


Fig. 4. Flowchart of Rules for Adjustment Mechanism.

the drone upward, z-axis. After the drone reached a certain desired height, it rotated at a yaw angle until desired tolerance error range is obtained. Then it moved to its side direction of the y-axis. The last stage is the forward direction, the x-axis. Whenever the drone carries out the defined tolerance in all the stages, the drone will be in the desired position. Algorithm 1 exhibits the pseudo-code, which is the scenario of the navigation system.  $\mu$  is the error tolerance.

**Algorithm 1** Pseudo Code of Navigation Algorithm

- 1: Start
- 2: Take off
- 3: Navigate drone in z-axis
- 4: if  $e_z < \mu_z$  :
- 5:   Navigate drone in yaw-axis
- 6: if  $e_z < \mu_z$  &  $e_{yaw} < \mu_{yaw}$  :
- 7:   Navigate drone in y-direction
- 8: if  $e_z < \mu_z$  &  $e_{yaw} < \mu_{yaw}$  &  $e_y < \mu_y$  :
- 9:   Navigate drone in x-axis
- 10: if  $e_z < \mu_z$  &  $e_y < \mu_y$  &  $e_{yaw} < \mu_{yaw}$  &  $e_x < \mu_x$  :
- 11:   Land
- 12: End

In the scenario of the navigation system, the trajectory is presented to lead the drone to reach its target point. Therefore, a desired path trajectory to the target can be predicted to avoid potential crashes. In addition, the velocity of the drone is determined as the angular velocity of the propellers that are dependent on body frame velocity,  $V^B$ .

**IV. RESULTS AND DISCUSSION**

In order to validate the navigation system, a model of the drone was created in a virtual environment called Gazebo integrated with Robot Operating System (ROS) [26], [27], [28].

To validate the performance of the tag-based navigation algorithm, a multi-target navigation strategy is utilized to move

the drone from the home point to the destination point. The multi-target navigation strategy is divided into three states. Each stage is marked by a tag with a different ID and the drone hovers in front of them before moving to the next stage. When the drone hovers in front of the first tag marker with the same procedure of hovering navigation, it rotates toward the next marker tag to face the next tag marker. This trend is followed for the next stages with different tag marker ID until the drone reach and land at the destination point. Therefore, the drone follows the defined trajectory to reach the target point without demanding the saved map in indoor/outdoor environment. Algorithm 2 represents the pseudo code of multi-target navigation.

**Algorithm 2** Pseudo Code of Multi-Target Navigation Algorithm

- 1: Start
- 2: Take off
- 3: Go forward to tag marker #1
- 4: Hover in front of tag marker #1
- 5: If  $e_x \leq \zeta_x$  and  $e_y \leq \zeta_y$  and  $e_z \leq \zeta_z$  and  $e_\phi \leq \zeta_\phi$  :
- 6:   Rotate  $90^\circ$
- 7: If  $e_\phi \leq \zeta_\phi$  :
- 8:   Go forward to tag marker #2
- 9: If  $e_x \leq \zeta_x$  and  $e_y \leq \zeta_y$  and  $e_z \leq \zeta_z$  and  $e_\phi \leq \zeta_\phi$  :
- 10:   Rotate  $-90^\circ$
- 11: If  $e_\phi \leq \zeta_\phi$  :
- 12:   Go forward to tag marker #3
- 13: If  $e_x \leq \zeta_x$  and  $e_y \leq \zeta_y$  and  $e_z \leq \zeta_z$  and  $e_\phi \leq \zeta_\phi$  :
- 14:   Land
- 15: End

Fig. 5 represents the comparison of the proposed navigation systems powered by APID and convention PID controller with constant parameters.

The results show that the navigation system enriched with APID converged faster to the target point significantly. Because it uses multiple sets of parameters for controlling based on the error range. When the error is in the F range, the drone moves

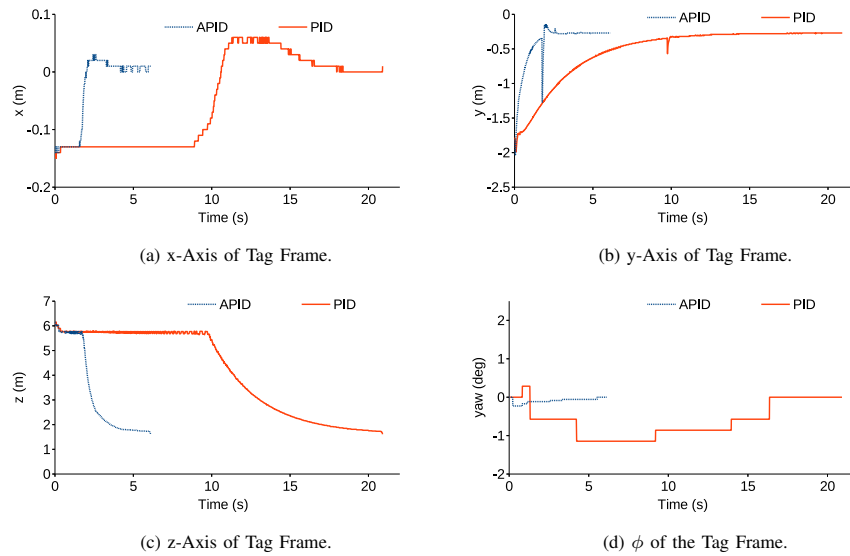


Fig. 5. Comparison of APID and PID Performance.

faster to the target than it is in N and AN ranges.

Fig. 6 shows the trajectory of the drone in  $x$  and  $y$  direction from the home point to the target point. The drone took off in the home point moved forward to tag #1, hovered in front of it, then rotated to the second tag. Similarly, the drone moved to the destination point.

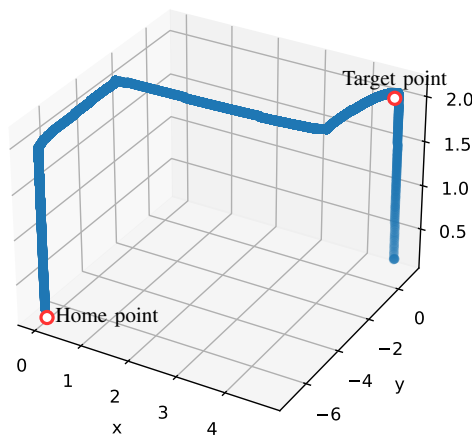


Fig. 6. Multi-Target Navigation.

The trajectory error and APID parameters changes are represented in Fig. 7 and 8, respectively.

In Fig. 8, it is represented that if the error is in the range of each set, the adjustment mechanism sets the set of parameters for the APID controller. For instance, in Fig.

8a between starting and 2.6s the APID parameters are the first parameters set,  $\Theta_{s3}$ , because the error is more than 1m. Subsequently, when the error is between 0.2m and 1m, from 2.6s to 3.16s,  $\Theta_{s2}$  is set. Finally, after 3.16s until the drone reached the target, the APID parameters follow  $\Theta_{s1}$ . In Fig. 8b, the switching time for APID parameter sets are 0.36 s and 1.57s. Furthermore, in Fig. 8c, at 4.46s the APID parameters turned from  $\Theta_{s3}$  to  $\Theta_{s2}$ , and it turned to  $\Theta_{s1}$  at 4.94s, when the error reduced from more than 1m to less than 1m and then decreased to less than 0.2m, respectively. In Fig. 8d, the turning times for APID parameters are 0.8s and 2.7s based on the ranges of the error shown in Fig. 7d. By considering the turning times, it can be obvious that turning times are in ascending order. The first turning times are 0.39s, 0.8s, 2.64 s, and 4.46s, for  $y$ , yaw,  $x$ , and  $z$  direction, respectively. Similarly, the second turning times are 1.57s, 2.7s, 3.16s, and 4.46s, for the  $y$ , yaw,  $x$ , and  $z$  direction, respectively. This mimics the scenario of the navigation system as shown in Algorithm 1.

## V. CONCLUSION

The paper presented a visual navigation system for an autonomous drone that hovers and navigates by using tag marker position and orientation autonomously. The kinematic and dynamic models of a drone have been determined by Newton- Euler. Moreover, AprilTag marker tag was introduced to obtain the position and orientation of the drone.

The results showed the navigation system using APID controller performed four times faster than the navigation system powered by the conventional PID controller. In addition, it demonstrated higher accuracy and reliable performance of the algorithms to accomplish various designed trajectories.

Besides its reliability and accuracy, one of the APID's advantages is less computational time rather than complex and memory consumable controllers. In addition, the tag navigation system and APID utilized an onboard camera available and usable for the light drones that cannot carry heavy sensors

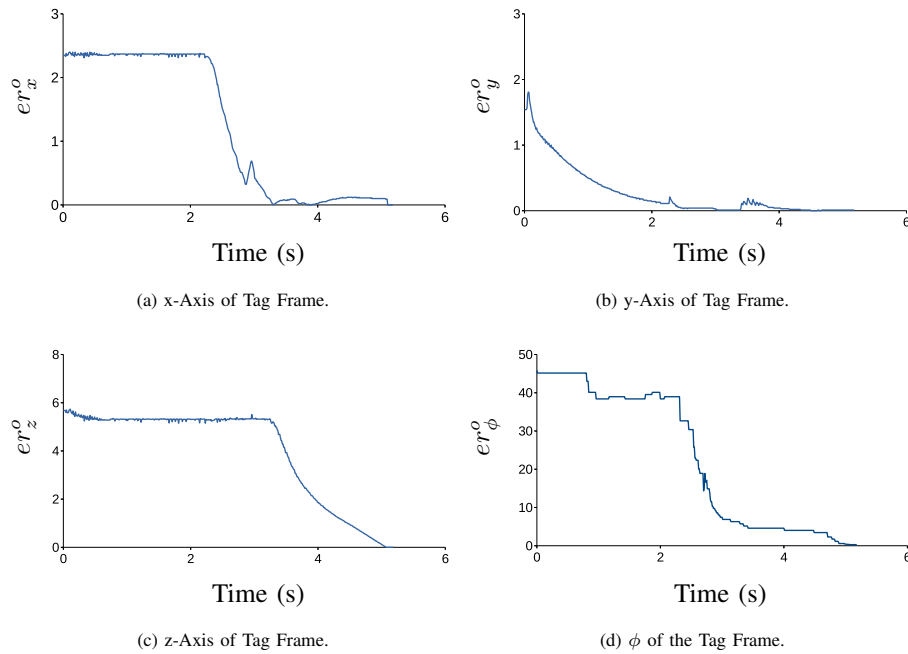


Fig. 7. Trajectory Error based on the Camera Frame.

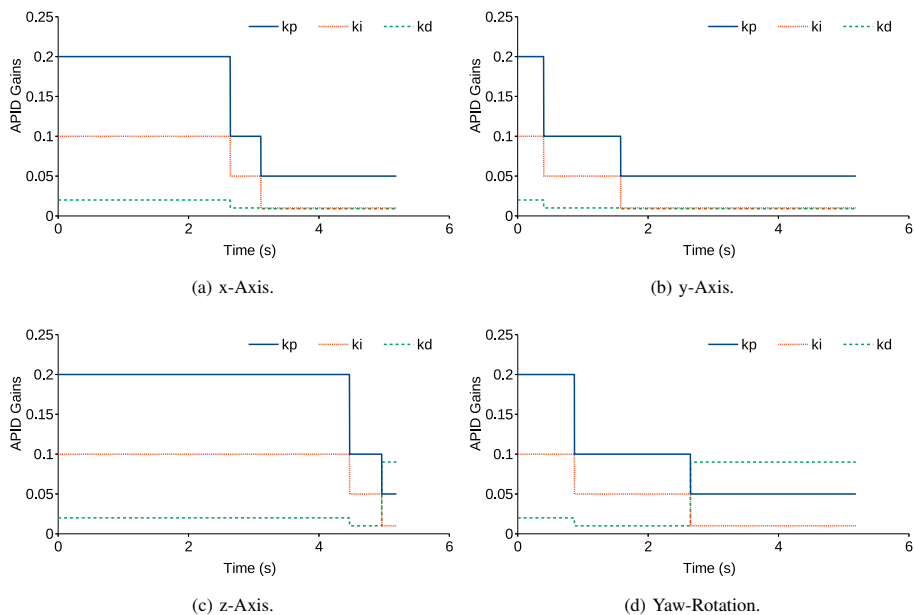


Fig. 8. Changes of the APID Controller's Parameters.

such as laser sensor for navigation and localization. Although it can be ascertained that the proposed method is efficient, there are some limitations. For instance, in an outdoor environment, the shininess of the tag marker can affect the view of the camera and navigation system. Moreover, the tag marker should be in the camera view for the drone to follow the desired trajectory. In future work, fuzzy and adaptive control systems can be investigated for visual-based navigation. In addition, this proposed algorithm can be validated for drone

navigation based on a laser sensor.

#### ACKNOWLEDGMENT

The authors would like to thank Universiti Kebangsaan Malaysia (UKM) and the Ministry of Higher Education Malaysia for the financial support received under research grant KK-2020-014 and TAP-K014062.



REFERENCES

- [1] C. Alex and A. Vijaychandra, "Autonomous cloud based drone system for disaster response and mitigation," *2016 International Conference on Robotics and Automation for Humanitarian Applications (RAHA)*, pp. 1–4, 2016.
- [2] A. Sagitov, K. Shabalina, L. Sabirova, and H. L. and Evgeni Magid, "Artag, apriltag and caltag fiducial marker systems: Comparison in a presence of partial marker occlusion and rotation," *Proceedings of the 14th International Conference on Informatics in Control, Automation and Robotics*, vol. 2, pp. 192–191, 2017.
- [3] P.-J. Bristeau, F. Callou, D. Vissière, and N. Petit, "The navigation and control technology inside the ar.drone micro uav," *IFAC Proceedings Volumes*, vol. 44, no. 1, pp. 1477–1484, 2011.
- [4] Y. Ham, K. K. Han, J. Lin, and M. Golparvar-Fard, "Visual monitoring of civil infrastructure systems via camera-equipped unmanned aerial vehicles (uavs): a review of related works," *Visualization in Engineering*, vol. 4, no. 1, pp. 1–8, 2016.
- [5] R. Casado and A. Bermudez, "A simulation framework for developing autonomous drone navigation systems," *Electronics*, vol. 10, no. 1, p. 7, 2021.
- [6] A. T. Azar, A. Koubaa, N. A. Mohamed, H. A. Ibrahim, Z. F. Ibrahim, M. Kazim, A. Ammar, B. Benjdira, A. M. Khamis, I. A. Hameed, and G. Casalino, "Drone deep reinforcement learning: A review," *Electronics*, vol. 10, no. 9, p. 999, 2021.
- [7] Y.-P. Huang, L. Sithole, and T.-T. Lee, "Structure from motion technique for scene detection using autonomous drone navigation," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 49, no. 12, pp. 2559–2570, 2019.
- [8] E. Cetin, C. Barrado, G. Munoz, M. Macias, and E. Pastor, "Drone navigation and avoidance of obstacles through deep reinforcement learning," *IEEE/AIAA 38th Digital Avionics Systems Conference (DASC)*, pp. 1–7, 2019.
- [9] X. Li and X. Yang, "Stability analysis for nonlinear systems with state-dependent state delay," *Automatica*, vol. 112, p. 108674, 2020.
- [10] S. Zahran, A. Moussa, and N. El-Sheimy, "Enhanced drone navigation in gnss denied environment using vdm and hall effect sensor," *ISPRS International Journal of Geo-Information*, vol. 8, no. 4, p. 169, 2019.
- [11] J. Upadhyay, A. Rawat, and D. Deb, "Multiple drone navigation and formation using selective target tracking-based computer vision," *Electronics*, vol. 10, no. 17, p. 2125, 2021.
- [12] V. Hodge, R. Hawkins, and R. Alexander, "Deep reinforcement learning for drone navigation using sensor data," *Neural Comput & Applic*, vol. 33, pp. 2015–2033, 2021.
- [13] V. R. F. Miranda, A. M. C. Rezende, T. L. Rocha, H. Azpurua, L. C. A. Pimenta, and G. M. Freitas, "Autonomous navigation system for a delivery drone," *J Control Autom Electr Syst*, vol. 33, pp. 141–155, 2022.
- [14] D. Tang, T. Hu, L. Shen, Z. Ma, and C. Pan, "Apriltag array-aided extrinsic calibration of camera–laser multi-sensor system," *Robotics and Biomimetics*, vol. 3, no. 1, 2016.
- [15] D. Malyuta, C. Brommer, D. Hentzen, T. Stastny, R. Siegart, and B. Roland, "Long-duration fully autonomous operation of rotorcraft unmanned aerial systems for remote-sensing data acquisition," *Journal of Field Robotics*, vol. 37, no. 1, pp. 137–157, 2020.
- [16] T. Lee, S. McKeever, and J. Courtney, "Flying free: A research overview of deep learning in drone navigation autonomy," *Drones*, vol. 5, no. 2, p. 52, 2021.
- [17] H. Yang, Y. Lee, S. Y. Jeon, and D. Lee, "Multi-rotor drone tutorial: systems, mechanics, control and state estimation," *Intelligent Service Robotics*, vol. 10, no. 2, pp. 79–93, 2017.
- [18] A. Ajami, J.-P. Gauthier, and L. Sacchelli, "Dynamic output stabilization of control systems: An unobservable kinematic drone model," *Automatica*, vol. 125, p. 109383, 2021.
- [19] K. V. Rao and A. T. Mathew, "Dynamic modeling and control of a hexacopter using pid and back stepping controllers," *4th International Conference on Power, Signals, Control and Computation*, pp. 1–7, 2018.
- [20] M. S. Amiri, R. Ramli, and M. F. Ibrahim, "Initialized model reference adaptive control for lower limb exoskeleton," *IEEE Access*, vol. 7, pp. 167 210–167 220, 2019.
- [21] C. Conker and M. K. Baltacioglu, "Fuzzy self-adaptive pid control technique for driving hho dry cell systems," *International Journal of Hydrogen Energy*, vol. 45, no. 49, pp. 26 059–26 069, 2020.
- [22] J. J. Castillo-Zamora, K. A. Camarillo-Gomez, G. I. Perez-Soto, and J. Rodriguez-Resendiz, "Comparison of pd, PID and sliding-mode position controllers for v-tail quadcopter stability," *IEEE Access*, vol. 6, pp. 38 086–38 096, 2018.
- [23] M. S. Amiri, R. Ramli, and M. F. Ibrahim, "Genetically optimized parameter estimation of mathematical model for multi-joints hip–knee exoskeleton," *Robotics and Autonomous Systems*, vol. 125, p. 103425, 2020.
- [24] D. Lee, S. J. Lee, and S. C. Yim, "Reinforcement learning-based adaptive pid controller for dps," *Ocean Engineering*, vol. 216, p. 108053, 2020.
- [25] S. M. Abbas, S. Aslam, K. Berns, and A. Muhammad, "Analysis and improvements in apriltag based state estimation," *Sensors*, vol. 19, no. 24, pp. 1–32, 2019.
- [26] M. S. Amiri and R. Ramli, "Intelligent trajectory tracking behavior of a multi-joint robotic arm via genetic–swarm optimization for the inverse kinematic solution," *Sensors*, vol. 21, no. 9, p. 3171, 2021.
- [27] M. S. Amiri, R. Ramli, M. F. Ibrahim, D. A. Wahab, and N. Aliman, "Adaptive Particle Swarm Optimization of PID Gain Tuning for Lower-Limb Human Exoskeleton in Virtual Environment," *Mathematics*, vol. 8, no. 11, p. 2040, 2020.
- [28] K. A. Juhari, R. Ramli, S. M. Haris, Z. Ibrahim, and A. Z. Mohamed, "Development of Floor Mapping Mobile Robot Algorithm Using Enhanced Artificial Neuro-Based SLAM (ANBS)," *Jurnal Kejuruteraan*, vol. 3, no. 1, pp. 59–64, 2020.

# Design of a Mobile Application for the Logistics Process of a Fire Company

Luis Enrique Parra Aquije<sup>1</sup>, Luis Gustavo Vásquez Carranza<sup>2</sup>, Gustavo Bernnet Alfaro Peña<sup>3</sup>,  
Michael Cabanillas-Carbonell<sup>4</sup>, Laberiano Andrade-Arenas<sup>5</sup>  
Facultad de Ingeniería y Negocios, Universidad Privada Norbert Wiener, Lima, Perú<sup>1,2,3,5</sup>  
Facultad de Ingeniería, Universidad Privada del Norte, Lima, Perú<sup>4</sup>

**Abstract**—Currently, the logistics process is an important part for any company because it helps to manage the assets and products that enter and leave it. Some companies carry out this process physically, saving the information on sheets of paper or Excel files, which takes longer to do and is not at the forefront of how companies do it, which is by using mobile applications to improve this process. Likewise, it has been decided to implement a mobile application with the aim of improving the logistics process in the Callao No. 15 fire company. For the elaboration of the application, the RUP methodology was used to do it in a more optimal way, in the end, a survey of experts in Google Forms was conducted, addressed to 10 experts to know the evaluation of the mobile application. In the end, a favorable result was obtained from the opinion of the experts on the mobile application; 70% of the respondents indicate that the usability of the mobile application has a “Very high” level; it can be seen that 80% of respondents indicate that the presentation of the mobile the application has a “Very high” level; it can be seen that 90% of the respondents indicate that the functionality of the mobile application has a “Very high” level; besides, it can be seen that 80% of the respondents indicate that the security of the mobile application has a “Very high” level.

**Keywords**—Fire company; logistics process; mobile application; RUP methodology; expert survey

## I. INTRODUCTION

From previous years to the present, in the fire company No. 15 of Callao they do not adequately apply the processes to be able to have an order in the logistics process. We found the inventories carried out by the firefighters themselves, since they have not defined how to carry out an entry and exit process in an inventory in the general services area. At the time of carrying out the inventory, the person in charge uses paper and a pen, in which he indicates the assets that belong to the company, and then passes it digitally to a program to be able to save, said program is called Microsoft Office Excel Spreadsheet. Although it is true that papers are used, this happens many times that these papers are lost, in the course of being digitized to the computer, causing loss of time, since everything has to be done again, which generates discomfort and dissatisfaction for part of the heads of the company and the areas, since they all work with the window to be able to be guided by where each element, object or object is located, since the company's warehouse is small, so they are also stored in other environments. In the program used in the company, when more than 500 articles are entered, the program is slow, since the computer is not prepared for this storage capacity. It

is also observed that everything is disordered in the program, showing a disorder by areas.

According to the authors Rasheed et al. [1], indicate that the World Health Organization (WHO) has decided to optimize its supply chains so that losses are lower in the medicines it has. In addition, the solution that was applied for this case was the development of a logistics software to control the medicines in its warehouse. Although the WHO in 2005 had already used software for this case, they decided to use more updated software. On the other hand, this new software has facilitated the logistics of medicines due to how friendly the system is, being faster, more efficient and safer.

On the other hand, the authors Bernal et al. [2], tells us The recommended solutions meet current and emerging needs in an agile B2B or B2C e-commerce environment. Rather, suppliers and customers offer and demand mobility, storage and customs services for the products it offers. In this case, your goal is to ensure that your logistics processes have added value from the beginning to the end of the business processes in the trade. On the other hand, this will also generate more profit for the companies that use them.

According to Pekarckikoba [3], when today's software and digital tools are used, the result is more efficient, taking less time to execute and giving a more optimal result. Furthermore, these tools have standards, are transparent and reduce unnecessary work. Rather, the modeling that was implemented and the e-kanban test when it was commercialized increased the efficiency of logistics in its processes. On the other hand, this model can be used as a digital solution for business processes.

Besides, the authors Fedorko et al. [4], suggest that the use of for the logistics of the company optimizes the storage of the products it offers in the warehouses. The objective is to reduce the time needed to collect the orders that the company's customers have made. Although the logistics processes have been optimized to cover various needs of the production processes in the warehouse. On the other hand, with the optimization of logistics, a time saving of 48 minutes and 36 seconds has been achieved.

The author Zhao [5], the goal is to find key points that affect customer satisfaction and warehouse utilization to achieve time reduction and minimize system congestion. In addition, a method called Box-Behnken was used to optimize and increase the speed of classification, distribution, storage through software. However, this method has parameters such as the speed of classification, distribution and storage capacity to

optimize the logistics of companies. Rather, when this method was used, a better response was obtained in the company's logistics system in terms of distribution and storage, achieving it in less time.

This project is carried out for the Callao No. 15 fire company, which will be implemented in the general services area, who is in charge of the logistics process, in this case inventories. This project is being carried out in order to optimize time and so that physical paper cannot be used, in which case it could be lost in the process of typing it into a computer. It will be done through a system, with the URP methodology which helps us to carry out the designs, programming, training, etc. This project is important, since with this you can improve the fire company, in the case of making everything more didactic and easy on the subject of inventories and thus make a digital transformation. The problem that is being solved is the losses and delays caused by doing everything on physical paper. The positive repercussions for this project is that they can be motivated to continue with the digital transformation in the different areas of the fire company. The negative repercussions of this project is that there will be people who will not want to adapt to new technologies and stay with the old as they are used to.

The people who are going to benefit from this project are the chief of general services, the deputies, and the first chiefs of the fire company, as well as all the firefighters, since they could also use it throughout their firefighter career.

The objective of this project is to facilitate logistics process for the fire company Callao No. 15, ordering the products that are used, the products that arrive, with the detail data of each product and secure the data through a mobile application design.

The article has the following structure: In Section II the literature review, in Section III establishes the Rup methodology, in Section IV the results, in Section V the discussions and finally Section VI with the conclusions and work future

## II. LITERATURE REVIEW

In this section of the article, emphasis was placed on the analysis of the research carried out on the subject of logistics software, where it will be seen what methods have been used and what their results were.

The authors Angolia and Pagliari [6], in this case a simulation is described in the dynamic environment for logistics decision-making with software. In addition, strategies for sales and operations were used to improve logistics and the supply chain in the company's inventory. On the contrary, this software helps to comply with the weight regulations that the products must have for the carriers. On the other hand, logistics software generates a great advantage for the recruitment of employees by the company.

The authors Byun et al. [7], take into account that the use of mobile applications in the logistics field called logistics in life. In addition, this sector has been used for mobile applications, which outperform traditional logistics companies. Likewise, the goal of this research is to measure whether these apps comply with the rules of use to satisfy users through measurement and evaluation. Specifically, evaluations of the

use of apps in Korea and other countries were carried out through the analysis of Big Data.

According to the author Nuanmeesri [8], this work has developed a mobile supply chain application for a window with a goal of marketing, distribution of products and logistics in agriculture and consumption according to the economic guide of Thailand 4.0. In addition, this work on the delivery supply chain of agricultural products focuses on trade, delivery and logistics using algorithms to find delivery routes for the products. Also this app fits the distribution of products, supply chain in an economy that is based on values. As a result, the mobile application has resulted in its effective use for the distribution of agricultural products in the supply chain in trade, distribution and logistics regarding Thailand 4.0.

The estimation of data on the reuse of waste by means of special statistical software and using the principles of logistics. Furthermore, such research goes beyond the research called "Environmental Assessment of Waste Recycling Based on Logistics Principles and Computer Simulation Design", which creates a succession of data that must be examined and evaluated separately. Likewise, these data that symbolize 15 classes of waste for 5 years, enter the analysis. As a result, the classes of waste that form a large part of the final manufacture of waste were located through descriptive statistics [9].

According to authors Ahmad and Bamnote [10], it informs us that software cost estimation (SCE) is an emerging concern for software companies during the software development period, because it asks for elements of effort and cost to create the software. In addition, these elements are created using Artificial Intelligence models, which can be less accurate and less reliable by increasing the risk factor of software projects. Said investigation. Also, this research proposed an algorithm whose function is to develop a model to optimize the price of the software. Since this research used a set of data collected from a software engineering database to run an effective performance study.

According to the authors Naseer et al. [11], tells us that Software engineering is a professional field in terms of education and practice. Software projects are key elements of software engineering courses. The goal of this project is to forecast the teams that are expected to achieve an evaluation that is not above average in the production of software products. The proposed method that stands out among the others in the forecast of teams that have a low performance in a phase of premature appreciation. The proposed method supported by J48 stands out from others by conceiving 89% correct forecasts.

The authors Garnov et al. [12], the reason for this research is to evaluate the effects of the digitalization method of cargo and logistics provision in the agriculture of the country of Russia. In addition, a technique has been used that handles information from Rosstat, GooglePlay provision and Yandex.Radar provision as information source. Likewise, this work reveals that one of the important trends in the field of freight transport is the use of technologies, such as transport and warehouse management systems, to automate the trading methods of carriers with mobile applications for shipments. and product orders. In conclusion, this digital system will connect all market users and increase the transparency and

traceability of freight transport.

The authors Rajabizadeh and Rezghi [13], the automated search for snake portraits will be able to help avoid venomous snakes and also offer an excellent method for patients. Also, in this work, k-nearest neighbors, support vector machine and logistic regression techniques are used in combination with principal component analysis and linear discriminant analysis as the feature extractor. Furthermore, this research shows MobileNetV2 as an effective deep neural network algorithm for snake portrait categorization that you can use even on mobile devices. Ultimately, this discovery shortens the path to creating mobile applications for the search for snake portraits.

Social networks today play a very important role in disaster response by locating needs in a short period of time and, therefore, optimizing situational awareness. In addition, this research aims to increase interest in a reserved search book to qualify this feature. Also, an overview chart is shown to check if it is feasible to act on the data created by the users before their review of the logistics planning of disaster replication. In conclusion, this research is adjustable to a diversity of logistics organization difficulties, this research shows its handling by means of a mobile delivery application of disaster replication basic goods [14].

The authors Barbosa et al. [15], the applied sciences of portraiture have increased to a significant horizon in recent years, used in different branches of research, such as those focused on the search for plants. Also, in this research, HerbApp, a mobile application that is used to separate herbal plants from non-herbal ones, is shown in order to spread the knowledge among people about the importance of plants. Likewise, other plant traits and traits are used to do pattern finding and data analysis. In conclusion, the experiences show that guidance facilitates efficient results.

In conclusion, to improve the logistics process of the Callao N°15 fire company, it is necessary to stop using Microsoft Excel as a database and start using a mobile application for this process, thus storing the data of the products that are used and the products that are used. The gap found is that there is still a lack of the use of emerging technologies that allow optimizing its processes.

### III. METHODOLOGY

In this section, the methodology that was used to develop and implement the mobile application for the Callao N°15 fire company, which is the object-oriented methodology called Rational Unified Process (RUP).

#### A. Processes of the RUP Methodology

According to the author Tia [16], RUP, as shown in Fig. 1, is a methodology for making software used by software companies. Likewise, there are different types of RUP software projects, which are small scale, large scale and re-engineering, which have needs and uses for each role. In addition, in this methodology there are different roles that one person cannot do because they work at the same time. In conclusion, the objective of the RUP methodology is to be able to develop high-quality software that meets the expectations of users and customers.

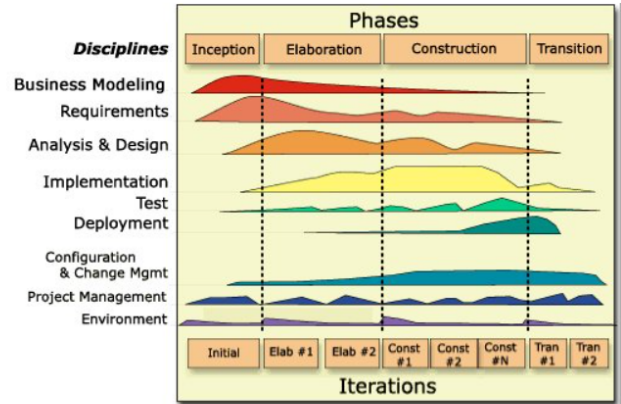


Fig. 1. Methodology RUP (Rational Unified Process).

In Fig. 2, it indicates that they have six processes (business modeling, requirements, analysis and design, implementation, test, deployment). It is a software engineering process that provides an orderly approach to the assignment of tasks and commitments in a software development company.

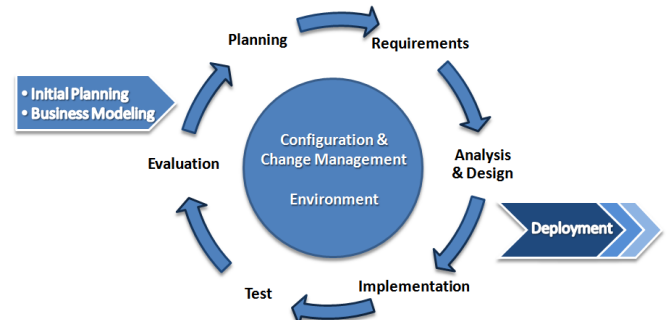


Fig. 2. Processes of the RUP Methodology.

#### B. Prototype Tools:

1) *Bizagi Modeler*: It is a Platform that consists of three components. Each of them facilitates a key step in the transformation and automation of your business processes. According to the author Germanía [17], Bizagi is a software for the representation of processes through sketches, structures, trade rules, actors. Likewise, the moments of the work process between their tasks are published, the time it takes for each task and the process are shown. Furthermore, the phases of Bizagi software development are modeling, building and process execution. In conclusion, the Bizagi software is very important when carrying out a process improvement [18], because it details what the process is like and what parts can improve to optimize the process.

2) *Star UML*: It is a software modeling tool based on the Unified Modeling Language (UML) and MDA (Model Driven Architecture) standards. The authors Naing et al. [19], Star UML is software for developing a fast, flexible, extensive,

full-featured and freely available Unified Modeling Language platform. It also uses UML for software design, uniting the notorious approaches to detailing organization and behavior. In addition, it is excellent in characterization for the user’s environment and has greater extensibility in its functions. In conclusion, Star UML is used to create detailed process diagrams in the UML language.

3) *Figma*: According to the author Arifin et al. [20], Figma is a prototyping interface layout program that runs on a web page via a browser. Likewise, said page layout program, but it is actually considerably more than that. In addition, this program can make software prototypes for computers, cell phones and tablets. In conclusion, the author’s final opinion about Figma is surely the best program for developing layout projects collaboratively as a team.

C. Development Tools:

1) *C#*: It is a component-oriented object-oriented programming language. According to the author Jankowski et al. [21], C# is an object-oriented, type-safe programming language. Likewise, it allows developers to establish several types of consistent and secure apps that run in the environment. In addition, it has its foundations in the C language group and will subsequently involve a group for C, C++, Java, and JavaScript developers. In conclusion, C# is a language that can be used for both desktop and mobile applications.

2) *Microsoft Visual Studio*: The Visual Studio IDE is a creative launch pad that you can use to edit, debug, and build code, and then publish an app. According to the author Hrabovskiy [22], Microsoft Visual Studio is an IDE (Integrated Development Environment) developed by the Microsoft company for software development. Likewise, it is available for OS (Operating Systems) such as Windows, Linux and macOS, not counting free and paid versions. In addition, it has compatibility with programming languages such as C++, C#, Visual Basic .NET, F#, Java, Python, Ruby, and PHP. In conclusion, Microsoft Visual Studio is an integrated development environment for making desktop, web, and mobile software.

3) *Microsoft SQL Server*: In this section was development of the six processes of the RUP methodology to know what steps the mobile application has followed to improve the inventory logistics of the Callao No. 15 fire company.

D. Development Methodology

In this section was development of the six processes of the RUP methodology to know what steps the mobile application has followed to improve the inventory logistics of the Callao No. 15 fire company.

In the Canvas model, it will be described how the mobile application is constituted in all aspects to know what objective it wants to achieve with its implementation.

1) *Requirements*: As shown in Table I and Table II, an interview was conducted with the client to find out the requirements that the mobile logistics application will have so that it works as the fire company wants. In the functional requirements table it will be shown what functions the logistics mobile application will have.

In the non-functional requirements table it will be shown how the logistics mobile application will work.

TABLE I. FUNCTIONAL REQUIREMENT

Non-functional requirement	Requirement Description	System Use Case	System Use Case Description
RF001	Register product	CUS001	Manage check-ins and check-outs
RF002	Search product	CUS001	Manage check-ins and check-outs
RF003	Modify product	CUS001	Manage check-ins and check-outs
RF004	Delete product	CUS001	Manage check-ins and check-outs
RF005	Consult product	CUS002	Operability of tools, accessories
RF006	Observe product status	CUS002	Operability of tools, accessories
RF007	Modify product status	CUS023	Operability of tools, accessories
RF008	Browse product	CUS003	Manage Location
RF009	Observe product location	CUS003	Manage Location

TABLE II. NON-FUNCTIONAL REQUIREMENT

Non-functional requirement	Requirement Description	Classification	Priority
RNF01	The response time for patient search of 8 seconds	Response Time and performance	High
RNF02	Access will be for authorized users only	Security	High
RNF03	User-friendly design and interface	Usability	High
RNF04	Navigability between fields and interface components (tab key, enter)	Usability	High
RNF05	A specific system function will be accessed based on your role	Security	High
RNF06	The application can support changes in its functions	Maintainability	High

2) *Analysis and Design*: As shown in Fig. 3 to Fig. 11, see how the logistics process of inventories works through the graphics made in Bizagi and StarUML, you will also see the design of the mobile application through a prototype.

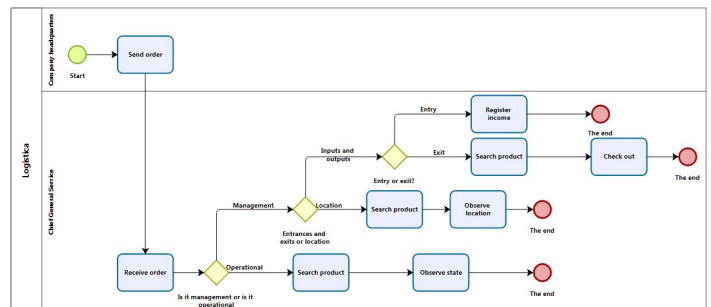


Fig. 3. Business Process Diagram.

In Fig. 3 is observed business process diagram will describe



in detail the logistics process of the business product company, which will describe how the process is carried out and who is involved in it.

In Fig. 4 shows the use case diagram of the system, it show

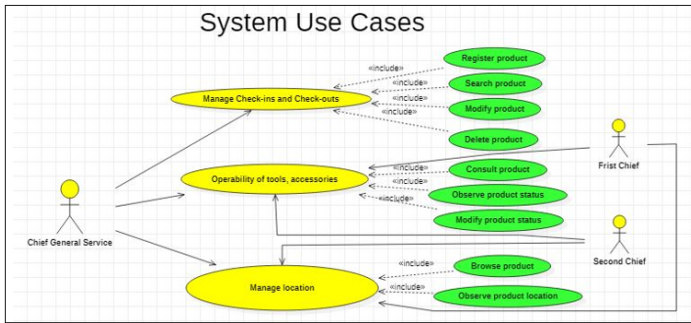


Fig. 4. System Use Case Diagram.

what functions the mobile application must have according to what the fire company requires and needs for its logistics process.

In Fig. 5 shows the input and output activity diagram, that

Manage Check-ins and Check-outs

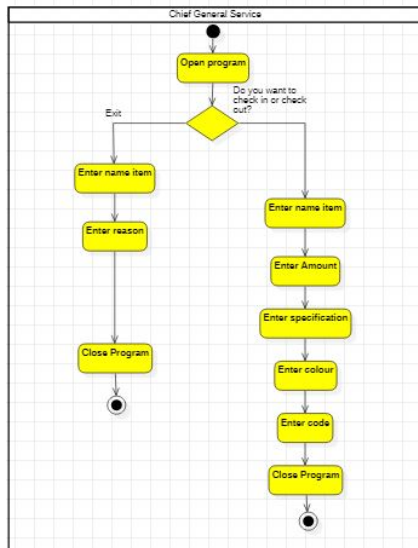


Fig. 5. Input-Output Activity Diagram.

the mobile application performs for the input or output of the products involved in this process.

In the operability activity diagram you will see the steps that the mobile application performs to see the outputs operability of the products involved in this process.

In the location activity diagram see the steps that the mobile application performs to see the location of the products involved in this process.

In the input sequence diagram, see how the user interacts with the mobile application when they input a product to the fire company.

In the output sequence diagram, see how the user interacts

with the mobile application when it outputs a fire company product.

In the sequence of operation diagram, look like the user interacts with the mobile application when needed to view the location of a fire company product.

In the operability sequence diagram it will be seen how the user interacts with the mobile application when he needs to see the location of a fire company product. In Fig. 6 shows

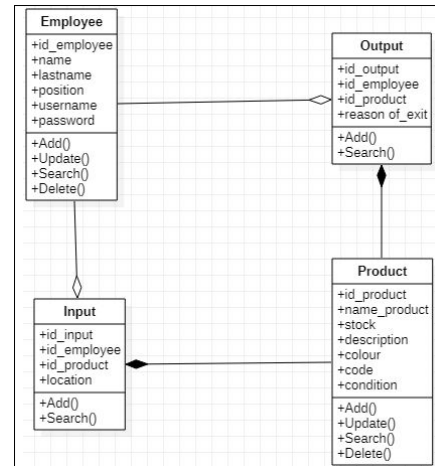


Fig. 6. Class Diagram.

the class diagram, it describes how the mobile application database is composed with its respective attributes and operations that each table contains.



(a) Star (b) Login

Fig. 7. Prototype Start and Login.

3) *Implementation:* The final design of the mobile application interfaces made in the Visual Studio IDE will be observed.

4) *Tests:* It will be observed how the inventory logistics mobile application works.





(a) New user (b) Menu

Fig. 8. Prototype New User and Menu.



(a) Outputs (b) Operability

Fig. 10. Prototype Outputs and Operability.



(a) Inputs - outputs (b) Inputs

Fig. 9. Prototype Inputs - Outputs and Inputs.

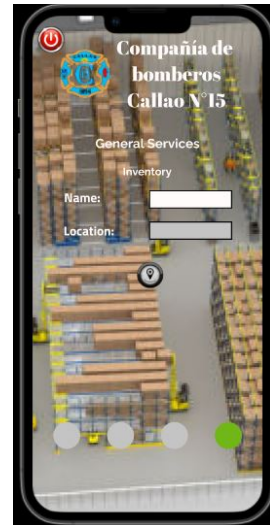


Fig. 11. Prototype Location.

5) *Deployment*: It was observed how the logistics mobile application will be executed on the mobile devices of the employees of the fire company.

#### IV. RESULTS

This section show the results of the case study, such as the RUP methodology, which was used in this work.

##### A. About Expert Judgment

As indicated in Table III, for this part a survey of experts was carried out in order to know how the experts feel with the implementation of the mobile application in the fire company,

dividing them into four criteria. which are usability, presentation, functionality and security. It was made with the likert scale from 1 to 5, where 1 is very low to 5 is very high.

The expert opinion table will show the survey that was done to each expert and what each one answered regarding the logistics mobile application.

1) *Usability Criteria*: Interpretation: From Table IV. It can be seen that 70% of the respondents indicate that the usability of the mobile application has a "Very high" level. In other words, there is 70% of those surveyed who reaffirm a very positive position and 30% consider it "High".

2) *Presentation Criteria*: Interpretation: From Table V. It can be seen that 80% of the respondents indicate that the presentation of the mobile application has a "Very high" level. In other words, there is 80% of those surveyed who reaffirm

TABLE III. EXPERT JUDGMENT

Criterion	Question	Half	Standart desviation	Scale
Usability	1. You feel that you are satisfied with the ease of use of the mobile application.	4.30	0.83	High
	2. You feel that you are comfortable with the use of the mobile application.	4.20	0.79	High
	3. You feel that the use of the mobile application is simple.	4.50	0.71	High
	4. You feel that using the mobile application you will complete your tasks in less time.	4.10	0.74	High
	5. You feel that using the mobile application you will complete your tasks more optimally.	4.50	0.71	High
Presentation	6. The mobile application interface is user friendly.	4.50	0.71	High
	7. The mobile app interface is well organized.	4.30	0.82	High
	8. The mobile application interface represents the image of the company.	4.30	0.82	High
	9. The colors of the mobile application interface have contrast with each other.	4.50	0.71	High
	10. The images of the interface help to understand what the functions of the mobile application are.	3.70	0.68	High
Functionality	11. The functions of the mobile application meet the needs of the user.	4.50	0.71	High
	12. Each functionality of the mobile application is properly located.	4.60	0.70	High
	13. The functions of the mobile application are understandable for the user.	4.00	0.94	High
	14. Mobile app features improve the way you get things done.	4.80	0.42	High
	15. The functions of the mobile application is adequate for this area.	4.50	0.53	High
Security	16. The mobile application is more secure allowing entry to only registered users.	4.50	0.53	High
	17. The mobile application protects the integrity of the information.	4.40	0.52	High
	18. The mobile application gives reliability to the information.	4.20	0.63	High
	19. The mobile application is safe for users.	4.60	0.52	High
	20. Mobile app security optimally protects data.	3.80	0.79	High

TABLE IV. USABILITY

		Usability (Bundled)			
		Frequency	Percentage	Valid percentage	Accumulated percentage
Valid	High	3	30,0	30,0	30,0
	Very high	7	70,0	70,0	100,0
	Total	10	100,0	100,0	

a very positive position and 20% consider it "High".

3) *Functionality Criteria:* Interpretation: From Table VI. It can be seen that 90% of the respondents indicate that the functionality of the mobile application has a "Very high" level. In other words, there is 90% of those surveyed who reaffirm a very positive position and 10% consider it "High".

4) *Security Criteria:* Interpretation: From Table VII. It can be seen that 80% of the respondents indicate that the security of the mobile application has a "Very high" level. In other words, there is 80% of those surveyed who reaffirm a very positive position and 20% consider it "High".

### B. About the Case Study

The mobile application has in its design 9 interfaces for its elaboration, operation and implementation for the fire company to improve its logistics process, which are:

- Star: It is the presentation of the software with the logo of those who carry it out [Fig. 7(a)].
- Login: It is the interface that will allow the registered user to enter a username and password registered in the application [see Fig. 7 (b)].
- New user: It is the interface that allows new users to register to have access to the mobile application [Fig. 8(a)].
- Menu: It is the interface that allows the user to see and access the functions of the mobile application [Fig. 8(b)].

- Inputs - outputs: It is the interface where it shows the functions that can be accessed and they are to input products and output them [see Fig. 9(a)].
- Inputs: It is the interface that allows the user to register a product through their data, which are names, description, stock, color, code and location see [Fig. 9(b)].
- Outputs: It is the interface that allows the user to exit the products that are no longer needed by means of their name and the reasons for which they are exiting [see Fig. 10(a)].
- Operability: It is the interface that allows the user to know and modify the operability of the product through its name [see Fig. 10(b)].
- Location: It is the interface that allows the user to know where it is located by means of its name (see Fig. 11).

### C. About the Methodology

- Benefits: The benefits of the RUP methodology are:
  - Encourages reuse of software code.
  - Reduce the difficulty of maintenance by making it easy to improve the changes that can be made and add more features to the software.
  - It makes it easy to reuse some functions of the software for other projects.
  - Maintenance is easier to perform using this methodology.
  - It can be applied to different software to make them of quality.

TABLE V. PRESENTATION

Presentation (Bundled)					
		Frequency	Percentage	Valid percentage	Accumulated percentage
Valid	High	2	20,0	20,0	20,0
	Very high	8	80,0	80,0	100,0
	Total	10	100,0	100,0	

TABLE VI. FUNCTIONALITY

Functionality (Bundled)					
		Frequency	Percentage	Valid percentage	Accumulated percentage
Valid	High	1	10,0	10,0	10,0
	Very high	9	90,0	90,0	100,0
	Total	10	100,0	100,0	

TABLE VII. SECURITY

Security (Bundled)					
		Frequency	Percentage	Valid percentage	Accumulated percentage
Valid	High	2	30,0	30,0	30,0
	Very high	8	70,0	70,0	100,0
	Total	10	100,0	100,0	

TABLE VIII. COMPARISON OF METHODOLOGIES

Advantage	RUP	SCRUM	XP
Documentation	5	4	3
Software projects	4	5	3
Presence in companies	4	5	3
Phases	4	5	3
Iteration	5	5	4
<b>Total</b>	22	24	16

- Comparison: As indicated in Table VIII, in this part you will make a comparison between the RUP, SCRUM and XP methodologies, this comparison will be made through 5 criteria that the methodologies have. The rating will be from 1 to 5, with 1 being very poor, 2 being poor, 3 being fair, 4 being good and 5 being excellent.

## V. DISCUSSIONS

The mobile application made in comparison to the prototypes the author Nuanmeesri [8], which is a mobile application for the logistics of agriculture and its consumption, whose function is used in the distribution of agricultural products to find routes for supply from the farms to its destination where it is marketed, instead the mobile application for the fire company apart from knowing the location of the products that enter and leave the fire company, you can know the state in which they are find the product and that can avoid mishaps when using it. Also, in the aspect of the function of the mobile application of the authors Garnov et al. [12], has the objective of managing the process of transport and warehouse logistics by automating this process through the use of technology for the processes of shipments and orders of products offered by Russian agriculture, a similar case to the one that has the fire department application that is being implemented for the logistics process for the fire company that also has the

same functions, but the information that is handled in the fire company is from its own database, otherwise it is the of the aforementioned authors who used information from other databases that are not their property but from other agencies or applications such as Rosstat, Google Play and Yandex.Radar, generating satisfaction in both cases for the users to whom it is addressed. Also take into account the security that you must have in the logistics part [23].

## VI. CONCLUSIONS AND FUTURE WORK

The conclusions of this article are that by using the mobile application it helped to improve the logistics process of the fire company, reducing the time in which said process is carried out and helping to organize the products that are inside the warehouse the company. Likewise, the RUP methodology is suitable for developing and implementing mobile applications, software and other technologies, related to logistics or sales of many companies that want to optimize and automate their processes to compete in the current market, these types of technologies being mandatory have the companies to grow more in the area in which it works. The expert judgment carried out for this mobile application has given an affirmative response to the implementation of the mobile application for logistics processes because its rating was always high or very high regarding the criteria in which the mobile application was rated. In conclusion, in the future, the RUP methodology should be used to create software related to logistics and sales, using mobile applications, software or other technologies to optimize company processes digitally, leaving behind the sheets and Excel files that they generate. an unnecessary waste of money and time for the company, preventing it from competing in the current market.

## REFERENCES

- [1] H. Rasheed, M. Usman, W. Ahmed, M. H. Bacha, A. Zafar, and K. S. Bukhari, "A shift from logistic software to service model: A case

- study of new service-driven-software for management of emergency supplies during disasters and emergency conditions by who," *Frontiers in Pharmacology*, vol. 10, 2019.
- [2] W. N. Bernal, M. A. Jimenez-Barros, D. J. Molineras, and C. D. Paternina-Arboleda, "Developing logistic software platforms: E-market place, a case study," vol. 11756 LNCS, 2019.
- [3] M. Pekarcikova, P. Trebuna, M. Kliment, M. Mizerak, and S. Kral, "Simulation testing of the e-kanban to increase the efficiency of logistics processes," *International Journal of Simulation Modelling*, vol. 20, 2021.
- [4] G. Fedorko, V. Molnár, and N. Mikušová, "The use of a simulation model for high-runner strategy implementation in warehouse logistics," *Sustainability (Switzerland)*, vol. 12, 2020.
- [5] N. Zhao and M. Wang, "Research on parameter optimization of the express warehousing and distribution system based on the box-behnen response surface methodology," *Advances in Civil Engineering*, vol. 2021, 2021.
- [6] M. G. Angolia and L. R. Pagliari, "Experiential learning for logistics and supply chain management using an sap erp software simulation," *Decision Sciences Journal of Innovative Education*, vol. 16, 2018.
- [7] D. H. Byun, H. N. Yang, and D. S. Chung, "Evaluation of mobile applications usability of logistics in life startups," *Sustainability (Switzerland)*, vol. 12, 2020.
- [8] S. Nuanmeesri, "Mobile application for the purpose of marketing, product distribution and location-based logistics for elderly farmers," *Applied Computing and Informatics*, 2019.
- [9] M. Straka, M. Taušová, A. Rosová, M. Cehlár, P. Kačmáry, M. Sisol, P. Ignác, and C. Farkas, "Big data analytics of a waste recycling simulation logistics system," *Polish Journal of Environmental Studies*, vol. 29, 2020.
- [10] S. W. Ahmad and G. R. Bamnote, "Whale-crow optimization (wco)-based optimal regression model for software cost estimation," *Sadhana - Academy Proceedings in Engineering Sciences*, vol. 44, 2019.
- [11] M. Naseer, W. Zhang, and W. Zhu, "Early prediction of a team performance in the initial assessment phases of a software project for sustainable software engineering education," *Sustainability (Switzerland)*, vol. 12, 2020.
- [12] A. P. Garnov, K. V. Ordov, I. O. Protsenko, N. A. Prodanova, V. Y. Garnova, and T. P. Danko, "Digitalization of transport and logistics services in russian agriculture," *LAPLAGE EM REVISTA*, vol. 7, 2021.
- [13] M. Rajabizadeh and M. Rezghi, "A comparative study on image-based snake identification using machine learning," *Scientific Reports*, vol. 11, 2021.
- [14] E. Kirac and A. B. Milburn, "A general framework for assessing the value of social data for disaster response logistics planning," *European Journal of Operational Research*, vol. 269, 2018.
- [15] J. B. Barbosa, V. I. Jabunan, T. A. K. Lacson, M. W. L. Mabayan, and G. M. M. Napone, "Herbapp: A mobile-based application for herbal leaf recognition using image processing and regularized logistic regression classifier," *International Journal of Innovative Science and Research Technology*, vol. 2, 2017.
- [16] T. K. Tia, "Simulation model for rational unified process (rup) software development life cycle," *SISTEMASI*, vol. 8, 2019.
- [17] A. V. A. Germania, "Workers assessment automation process through the bizagi platform," *RISTI - Revista Iberica de Sistemas e Tecnologias de Informacao*, vol. 2020, 2020.
- [18] A. Ramos-Romero, B. Garcia-Yataco, and L. Andrade-Arenas, "Mobile application design with iot for environmental pollution awareness," *International Journal of Advanced Computer Science and Applications*, vol. 12, 2021.
- [19] A. A. Naing, M. T. Nyo, and S. M. Han, "An analysis design for betel-nut business by using unified modeling language (uml)," *University Journal of Creativity and Innovative Research*, vol. 01, 2020.
- [20] Y. Arifin, E. P. Gunawan, and M. Ohyver, "Development web application for enhancing information and activity in rptra maya asri 13," *ICCD*, vol. 3, 2021.
- [21] M. Jankowski and M. Skublewska-Paszowska, "Analysis of the use of java and c# languages for building a mobile application for the android platform," *Journal of Computer Sciences Institute*, vol. 16, 2020.
- [22] Y. Hrabovskyi, N. Brynza, and O. Vilkhivska, "Development of information visualization methods for use in multimedia applications," *EUREKA, Physics and Engineering*, 2020.
- [23] F. Andrade-Chaico and L. Andrade-Arenas, "Projections on insecurity, unemployment and poverty and their consequences in lima's district san juan de lurigancho in the next 10 years," 2019.

# Intelligent System for Personalised Interventions and Early Drop-out Prediction in MOOCs

ALJ Zakaria

University Sidi Mohamed Ben Abdellah  
Faculty of Science Dhar-Mahraz  
Fez, Morocco

BOUAYAD Anas

University Sidi Mohamed Ben Abdellah  
Faculty of Science Dhar-Mahraz  
Fez, Morocco

Cherkaoui Malki Mohammed Ouçamah

University Sidi Mohamed Ben Abdellah  
Faculty of Science Dhar-Mahraz  
Fez, Morocco

**Abstract**—In this paper, we propose an approach to early detect students at high risk of drop-out in MOOC (Massive Open Online Course); we design personalised interventions to mitigate that risk. We apply Machine Learning (ML) algorithms and data mining techniques to a dataset extracted from XuetangX MOOC learning platforms and sourced from the KDD cup 2015. Since this dataset contains only raw student log activity records, we perform a hybrid feature selection and dimensionality reduction techniques to extract relevant features, and reduce models complexity and computation time. Besides, we built two models based on: Genetic Algorithms (GA) and Deep Learning (DL) with supervised learning methods. The obtained results, according to the accuracy and the AUC (Area Under Curve)-ROC (Receiver Operator Characteristic) metrics, prove the pertinence of the extracted features and encourage the use of the hybrid features selection. They also proved that GA and DL are outperforming the baseline algorithms used in related works. To assess the generalisation of the approach used in this work, The same process is performed to a second benchmark dataset extracted from the university MOOC. Then, a single web application hosted on the university server, produces an individual weekly drop-out probability, using time series data. It also proposes an approach to personalise and prioritise interventions for at-risk students according to the drop-out patterns.

**Keywords**—MOOC; drop-out; dimensionality reduction; features selection; personalised intervention

## I. INTRODUCTION

Following the emergence of new digital technologies aiming to modernize the traditional education system. Massive Open Online Courses (MOOCs) have gained popularity in recent years [1]. In 2020, 16,300 courses were offered by 950 universities, and the number of enrollment has reached more than 180 million learners worldwide [2]. MOOCs have become an ideal source of self development that bridges the gap between industry requirements and skills acquired in the university [3]. Despite these benefits that bring a substantial improvement to the student learning experience. MOOCs are facing many problems today. Among the most cited are the high drop-out and the low completion rate. The average completion rate for a MOOC is 12.6% [4]. We can also cite the weak interactions and the absence of tutor support to the significant number of enrolled participants. This excessive attrition rate in MOOCs, has prompted researchers to consider the use of learning analytic for early prediction of learners at risk of drop-out [5].

Learning analytic consists of analysing the log trace and the data collected while students interact with the MOOC courses

[6]. Then, using supervised machine learning, mathematical models can automatically detect student at-risk of drop-out based on their previous behaviour and interactions during the course. The use of learning analytic has shown an encouraging potential and reduced visibly the attrition rates in MOOCs [7], [8]. However, it is limited, given the enormous number of enrolment either by 1) the late detection of students at-risk, by 2) the absence of prioritization which will considerably reduce the number of students that the instructor must address each time. Finally 3) the prediction models provide no clues for the monitor to propose a personalized intervention for each droppers pattern. Thus, a system that is capable of detecting at-risk students and providing a customised intervention is therefore needed.

The aim of this paper is to build an intelligent system using DL and GA in the field of learning analytic, and able to overcome these three limitations. This paper will describe the process followed to address the problem related to early prediction student drop-out, prioritizing student needing intervention according to a weekly temporal model prediction based on student drop-out probability. In the light of the obtained results, a timely personalized intervention can be designed and delivered to retain students at-risk. The obtained system gives the possibility to meet the challenges of identifying on a large scale students at high risk of dropping- out, while also satisfying the requirement to be able to support early intervention.

This paper is organised as follows: the next section present a brief overview of related work. The Section 4 is devoted to describe the different components of the used dataset and the extracted features. Section 5, is dedicated to presenting the methods used to predict student drop-out. Section 6, is dedicated to discussing the obtained results. Finally, we conclude the paper and give some perspectives on future work.

## II. RELATED WORK

Many researchers have recently focused on MOOCs student drop-out. Time consideration is very significant when tackling this problem. Early detection plays a masterful role in reducing the attrition rate. In fact, several studies have proved that 75% of drop-outs occur in the first weeks [9]. Similarly, Gitinabard et al. [10] analyse students logs and forum data of an annual MOOC lessons. They apply Logistic Regression and Support Vector Machine (SVM) to predict drop-out in the first weeks of each course. The students were flagged as drop-out with the precision of 70% after the third

week. Berens et al. [11] developed an early predicting system using demographic data and a boosting algorithm combining several ML algorithms: Linear Regression, Neural Network, Decision Tree, AdaBoost. The system provides an accuracy of 58.2% for the first semester and 81.5% for the fourth semester. Authors in [12] used deep learning and achieved an accuracy of 0.92% in the first week. In [13] ALJ et al. used several baseline machine algorithms and obtain an AUC ROC score of 90%. The performance of machine learning algorithms highly depends on the used dataset used. ML algorithms are unable to provide effective parameter setting method. Therefore, feature selection of parameters is another research content of this paper. A hybrid features selection method is proposed. Besides, Genetic Algorithms (GA) are used for parameters tuning.

GA are generally used for optimization problems and tuning classifiers in multiple fields such as emotion recognition [14] and medicine [15], and feature selection [16]. Despite GA are not quite used in the field of student drop-out, they are producing significant results.

This study contributes to the current state-of-the-art of the field in two main directions. First, by developing a comprehensive approach for studying and detecting early drop-out in a data perspective. Second, by designing a prioritized and personalized intervention for student at-risk of drop-out.

### III. DATA

#### A. Data Description

This study use a dataset provided from KDD cup [17], an annual Data Mining and Knowledge Discovery competition organised by ACM Special Interest Group. In KDD cup 2015, the dataset used in the competition contains users trace log extracted from XuetangX which is one of the biggest Chinese MOOC learning platforms. The aim was to predict users drop-out using different data mining techniques and machine learning algorithms.

The detailed description of the five parts of dataset is as follows:

- 1) The first part of the dataset contains information about the start and the end date of each course according to the Table I.

TABLE I. COURSE INFORMATIONS

Fields	Type	Description
Course-ID	Nominal	Course Identifier
Course-S	Date	Course Starting Date
Course-E	Date	Course Ending Date

- 2) Information about the modules of each course, sub-modules and also the category of modules and their start date according to the Table II.

TABLE II. MODULE INFORMATIONS

Fields	Type	Description
Course-ID	Nominal	Course Identifier
Module-ID	Nominal	Module Identifier
Module-cat	Nominal	Module Category
Module-child	Nominal	Sub-Module
Module-S	Date	Module Starting Date

- 3) Informations about enrolments: an enrolment is a (Student, Course) entry according to the Table III.

TABLE III. ENROLMENT INFORMATION

Fields	Type	Description
Enrolment-ID	Nominal	Enrolment Identifier
Student-ID	Nominal	Student Identifier
Course-ID	Nominal	Course Identifier

- 4) The fourth part of the dataset contains a log trace of every enrolment, log timestamps, and the source and the type of the event according to the Table IV.

TABLE IV. LOG INFORMATION

Fields	Type	Description
Enrolment-ID	Nominal	Enrolment Identifier
Student-ID	Nominal	Student Identifier
Course-ID	Nominal	Course Identifier
Event-Time	Timestamps	Time when the event occurs
Source	Nominal	Source of Event (Server / Browser)
Event	Nominal	e1 Problem
		e2 Access
		e3 Video
		e4 Wiki
		e5 Discussion
		e6 Navigate
		e7 Page Close

- 5) The last part of the dataset contains information about the real value of enrolment result according to the Table V

TABLE V. ENROLMENT RESULT

Fields	Type	Description
Enrolment-ID	Nominal	Enrolment Identifier
Result	Boolean	0 Success
		1 Drop-out

The dataset captures a trace log of 79186 students and 120543 enrolment, because every student can enrol in multiple courses. If a user leaves no records for course C in the log during the next 10 days, it is defined as drop-out from the course.

We notice that for all courses the drop-out represented by 1 in the table of enrolment results exceeds 65%. The majority class is drop-out with (95581) 79% compared to success (24961) 21% of enrolments. In this case, a class imbalance problem is faced. In order to balance the dataset, we will oversample the minority class in order to increase its cardinality to be equal to the majority class.



Oversampling: this technique duplicates copies of some points from the minority class to increase its cardinality to be equal to the majority class. New samples can be generated by random repetition or using more sophisticated methods such as SMOTE or ADASYN.

Synthetic Minority Oversampling Technique (SMOTE) [18] uses the KNN algorithm to generate new synthetic data points that combine features of the data point and its K closest neighbours. However, it still has some weaknesses regarding the oversampling logic used. On the one hand, it does not consider generating new samples from neighbours which may come from the other class. The synthetic observations can overlap with other observations of the majority class. On the other hand, generating multiple synthetic observations risks introducing additional noise in the dataset, this could potentially bias the model.

ADASYN [19] for Adaptive Synthetic, a version of SMOTE that has been improved. Instead of generating the same number of synthetic observations for each observation of the minority class, ADASYN adapts the oversampling to the distribution density of the observations of the minority class. Concretely, it produces more synthetic samples in regions of feature space where the density of minority observations is low and fewer samples in regions with higher density.

The selection of the technique to use remains strictly linked to the data set used. For our example, SMOTE gives better results. Finally, we end up with two datasets Unbalanced Dataset (UB) and Balanced Dataset (BD).

### B. Feature Engineering

The information available on the KDD cup 2015 dataset lacks personal information (e.g. age, sex, nationality) and information regarding the course (e.g. prerequisites, difficulty level). The logging trace remains the most potent source of information. Our feature extraction method is based on counting the log of every enrolment; an enrolment is a (student, course) entry. The extracted features can be divided into two parts:

**Enrolment History Features :** It contains features about the history of interaction with the MOOC, such as the number of successful courses, the number of failed courses, the cumulative number of days spent on the MOOC during old registrations, and the cumulative number of logs of each event present on the catalogue in Table IV.

**Current Enrolment Features:** It contains features about the number of days spent on the MOOC during the current enrolment and the count of logs of each event.

We also extract the count of minutes spent for every enrolment; after examining the log trace, we notice that all sessions start with the **Navigate** event and sometimes with the **Access**. We calculate the difference of time expressed in minutes between one of the two events and the end of the session expressed with the **Page\_close**. The accumulation of minutes is recorded for each enrolment in the variable m according to the following algorithm:

### ALGORITHM 1

Algorithm of Connected Minutes

```
Data : Raw log data
Result: Connected minutes per enrolment
@ConnectedMin = 0; @BeginDay = "00:00:00"
@EndDay = "23:59:59"; @TBegin = ""; @TEnd = "";
while not at the end of enrolment log rows do
  read current
  if @Event='Navigate' or (@Event = 'Access' and
  @TimeBegin = "") then
    @TimeBegin= Time-Event;
  end if
  if @Event='Page close' then
    @TimeEnd= Time-Event;
  end if
  @Diff=DateDiff(MIN,@TBegin,@TEnd);
  if @Ddiff > 0 then
    @x=DateDiff(MIN,@hdebut,@EndDay);
    @y=DateDiff(MIN,@BeginDay,@hfin);
    @Ddiff= @x+@y;
  end if
  @ConnectedMin= ConnectedMin+@Ddiff;
end while
```

Finally we end up with features presented in the Table VI.

TABLE VI. EXTRACTED FEATURES

N	Features
1	Enrolment Identifier
2	Count of student previous enrolments
3	Count of student previous succeeded enrolments
4	Count of student previous drop-out enrolments
5	Count of log for the current enrolment
6	Count of log for all previous enrolments
7	Count of days between first and last log
8	Count of days between first and last log for all previous enrolments
9	Count of log for the event : Problem
10	Count of log for the event : Problem for all previous enrolments
11	Count of log for the event : Video
12	Count of log for the event : Video for all previous enrolments
13	Count of log for the event : Navigate
14	Count of log for the event : Navigate for all previous enrolments
15	Count of log for the event : Page-close
16	Count of log for the event : Page-close for all previous enrolments
17	Count of log for the event : Access
18	Count of log for the event : Access for all previous enrolments
19	Count of log for the event : Discussion
20	Count of log for the event : Discussion for all previous enrolments
21	Count of log for the event : Wiki
22	Count of log for the event : Wiki for all previous enrolments
23	Count of logs in the first 10 days of course
24	Count of logs in the second 10 days of course
25	Count of logs in the last 10 days of course
26	Count of active minutes
27	Count of active days
	Enrolment result : Success 0 /Drop-out 1

### C. Feature Selection

When solving a classification problem, processing extracted data vectors is an important step. Indeed, the performance of the classifier highly depends on the correct choice of the content of these vectors. However, the problem becomes difficult to re-solve and very expensive in terms of training time and resources owing to the large dimension of these vectors. Consequently, it is useful, and sometimes necessary to reduce the dimensionality of these vectors to be compatible with resolution methods, even if this reduction may lead to a slight loss of information. Reducing the number of explanatory variables has a double advantage. On one hand the model will be easily interpretable due to the few number of variables. On the other hand, the prediction error will be reduced by removing the non-informative variables.

Feature selection is a process allowing to select a subset of features considered as the most relevant from a starting set using various criteria and different methods. The process is working as follows:

- 1) From the initial set of variables, the selection process determines a subset of variables that he considers to be the most relevant.
- 2) The subset is then evaluated with the classifier to assess the performance and relevance of the selection.
- 3) Depending on the result of the evaluation, a criterion for stopping the process determines whether the subset of variables can be used in the learning process, otherwise another subset of variables is generated and tested. The stopping criteria can be a predefined number of features to keep or a fixed number of iterations or even a criterion related to the evaluation function.

Methods used for selection can be classified into three main categories: Filter, Wrapper and Embedded.

1) *Filter*: The filter approach was the first method for selecting features [20]. It is considered a pre-processing step before the learning phase; the evaluation of features is usually done independently of the classifier. We define an importance score for each feature that reflects its quality as a predictor. We also define a score of similarity between two characteristics. The objective is to select the variables with the highest importance score and the lowest similarity scores to reduce redundancy. The main advantage of filtering methods is their computational efficiency and robustness against overfitting. Unfortunately, these methods do not consider interactions between characteristics and tend to select characteristics involving redundant rather than complementary information. The filter method used in this work is the Chi-squared test.

2) *Wrappers*: The main drawback of filter approaches is ignoring the influence of the selected variables on the learning algorithm's performance. To solve this problem, Kohavi and John introduced the concept of a Wrapper for selecting features [21]. Wrappers use the accuracy of the learning algorithm as an evaluation function to estimate the relevance of the variable. The Wrapper methods are generally considered to be better than those of filtering. They can select proper small subsets of features. However, features selected by this method are only

suitable to the classification algorithm and are not necessarily valid if we change the classifier. Also, the complexity of the learning algorithm makes the Wrapper methods very expensive in terms of computation time. It has been demonstrated by [21] that Wrapper methods produce better performance than some filtering methods. This paper will use Recursive Feature Elimination with both Random Forest (RFE-RF) and Gradient Boosting (RFE-GB).

**Recursive Feature Elimination (RFE)** is a selection algorithm based on backward elimination, in which recursive elimination aim to select a subset of optimum features. The learning is performed first with all the  $p$  variables, the least discriminant variable is removed, then the learning is performed on the  $p-1$  remaining variables. This process is iterated until the number of desired variables is obtained.

3) *Embedded*: Embedded methods incorporate the selection of variables into the learning process. Embedded methods can use all the dataset as a training set which is an advantage that can improve the result. In addition, Guyon and Elisseeff [22] specify that Embedded approaches surpass Wrapper approaches concerning the computation times and the robustness against overfitting. In this study, we are using both regularization: Ridge and lasso as Embedded methods.

**Regularisation** in ML adds a penalty term to the different coefficients of the model. The main purpose of Regularization is to avoid overfitting by improving the generalisation of the model. It improves the performance of models on new data. the main types of regularisation are Ridge and lasso:

Lasso regression (L1): Adds the squared magnitude of coefficients as a penalty term to the loss function.

$$L + \lambda \sum_{i=0}^n x_i^2$$

Ridge regression (L2): Adds the absolute value of magnitude of coefficients as a penalty term to the loss function.

$$L + \lambda \sum_{i=0}^n |x_i|$$

It is therefore used for variable selection. While L1 set the coefficient of unnecessary variables to 0, L2 is approaching them to zero.

In this study, we will implement a hybrid approach that combines all the methods seen previously. Each method will participate to elect if the variable will be selected or not. The scores obtained for each method will be then aggregated and normalised so that they are between 0 for the lowest rank and 1 for the highest. Variables with a high average will be selected according to score obtained in the Table VII.

TABLE VII. FEATURES RANKING

N	Lasso	REF-RF	REF-GB	Ridge	Chi-2	R-Lasso	Mean
1	0.0	1.0	0.7	0.0	0.0	0.48	0.36
2	0.00	0.71	0.05	0.07	1.00	0.01	0.31
3	0.00	0.29	0.75	0.21	1.00	0.91	0.53
4	0.00	0.57	0.35	0.14	1.00	0.00	0.34
5	0.48	1.00	0.90	0.39	0.97	0.00	0.62
6	0.06	1.00	0.2	0.00	1.00	0.00	0.38
7	0.00	1.00	1.00	0.05	1.00	0.89	0.66
8	0.00	1.00	0.45	0.00	1.00	0.01	0.41
9	0.00	1.00	1.00	0.01	0.99	0.02	0.50
10	0.00	1.00	0.15	0.00	1.00	0.14	0.38
11	0.00	1.00	0.80	0.00	1.00	0.34	0.52
12	0.00	1.00	0.60	0.00	1.00	0.00	0.43
13	0.00	1.00	0.55	0.01	1.00	0.62	0.53
14	0.00	0.00	0.00	0.00	1.00	0.00	0.17
15	0.00	1.00	1.00	0.02	1.00	0.13	0.53
16	0.00	1.00	0.25	0.00	1.00	0.00	0.38
17	0.00	1.00	0.95	0.00	0.99	0.28	0.54
18	0.00	1.00	0.40	0.00	1.00	0.01	0.40
19	0.00	0.14	0.10	0.00	1.00	0.00	0.21
20	0.00	0.86	0.65	0.00	1.00	0.00	0.42
21	0.00	1.00	0.30	0.04	1.00	0.02	0.39
22	0.00	0.43	0.50	0.00	1.00	0.00	0.32
23	0.00	1.00	1.00	0.40	0.99	1.00	0.73
24	0.00	1.00	0.85	0.39	0.99	0.87	0.68
25	1.00	1.00	1.00	0.38	0.99	1.00	0.90
26	0.17	1.00	1.00	0.00	0.67	1.00	0.64
27	0.00	1.00	1.00	1.00	1.00	1.00	0.83

According to the results found on the Table VII. The dimensionality of the dataset will be reduced to the eight following features:

- Count of student previous succeeded enrolments
- Count of log for the current enrolment
- Count of days between first and last log
- Count of log for the event : Access
- Count of logs in the second 10 days of course
- Count of logs in the last 10 days of course
- Count of active minutes
- Count of active days

In the next section, in order to prove the relevance of our selection, results obtained with this set of features will be compared with the results obtained using the initial set of variables.

#### IV. METHODOLOGY

The variable enrolment result has two alternative outcomes 1 for drop-out and 0 for success. Thus, the problem can be modelled as a binary classification. Besides, since the data used for training and testing is already labelled, the models are built with supervised learning. Fig. 1 present the methodology used in this paper.

Supervised learning begins with the training process. During training, the algorithm optimises the mapping function through a pair consisting of an input vector and the desired output value. The goal is to create a model that correctly classifies new unseen data. The predicted outputs are then

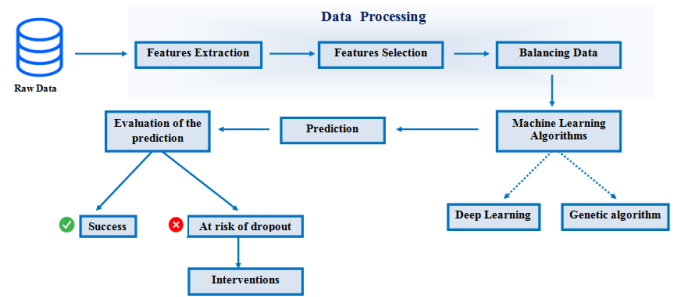


Fig. 1. Methodology.

compared to the accurate observation on the validation set to compute the model’s performance and generalisation ability. In this study, we will implement both GA and DL to create our models.

#### A. Deep Learning

DL is based on the models of NN with many hidden layers, called DNN. While a traditional NN can only handle a single hidden layer as show in the Fig. 2. DL data processing is carried through multiple layers to compute the output. Each layer is made of many artificial neurons imitating the biological neurons in a very simplified way. Each connection in the network is characterised by a coefficient or a synaptic weight that mainly describes the behaviour of the network.

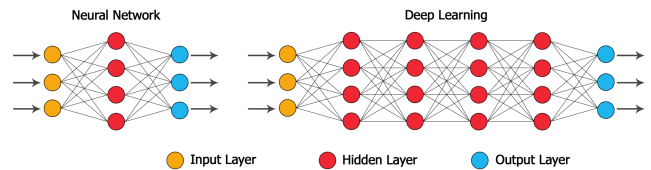


Fig. 2. Deep Learning.

During the learning process, weights are calculated in order to determine whether to amplify or dampen the output. The weights are adapted to minimise the difference between the network output and the expected output.

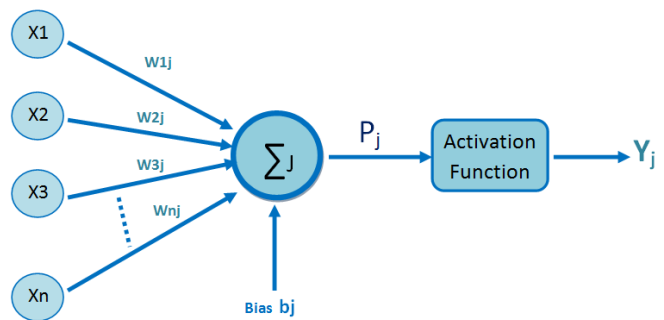


Fig. 3. Artificial Neuron.

As Fig. 3 shows, the neurons receive the information produced by other nodes through the input connections. Each

neuron J performs a weighted sum of the n input values. Weights assigned to the neuron’s inputs are stored in a matrix W. The value  $w_{ij}$  represents the weight of the input connection  $X_i$  of neuron j. Then to this sum a bias  $b_j$  is added. This total represents the biased post-synaptic potentials formulated as follows:

$$P_j = \sum_{i=1}^N w_{i,j} X_i + b_j$$

Finally, an activation function f transforms this biased potential to obtain the activation value of the neuron. This value is then transmitted to other neurons. Among the commonly used activation functions we can find Rectified linear unit (ReLU), Sigmoid and Hyperbolic tangent.

$$X_j = f(P_j)$$

### B. Genetic Algorithms

GA are stochastic optimization methods belonging to the family of evolutionary algorithms. They are commonly used for resolving complex optimization and search problems. GAs are inspired by the Darwinian mechanisms of the natural evolution of biological populations and rely on derived techniques such as selection, mutation and crossing. GA use the principle of survival of individuals considered to be the strongest or best suited to the environment by combining the strengths of each individual to create the next generation considered to be a better solution to the problem. This process is repeated several times until finding individuals have genetic information that corresponds to the best solution to the problem. In general, the process of a GA as presented in Fig. 4 is based on the following phases:

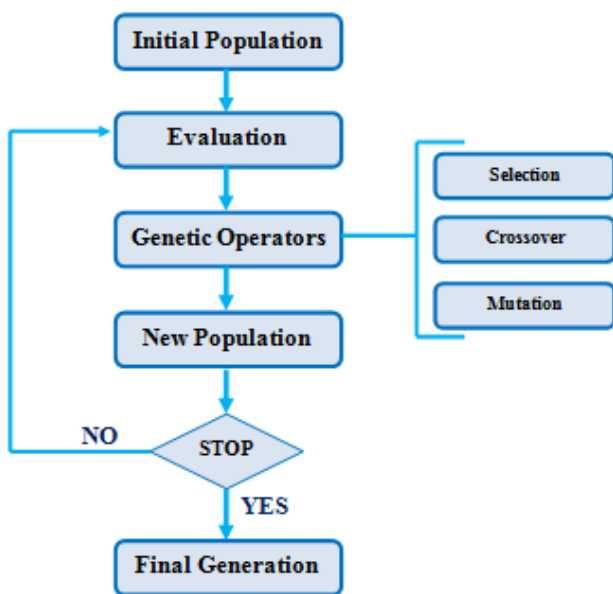


Fig. 4. Architecture of a Genetic Algorithm.

(1) **Initialization:** This GA mechanism must produce a non-homogeneous population of individuals who will serve as a parent for future generations. The choice of the initial population is important because it can make the convergence to the global optimum more or less rapid. The initial population must be distributed over the entire research area.

(2) **Evaluation:** Evaluate each capacity to the target variable with high accuracy. The LR algorithm was used to build prediction models. Individuals selected by the GA search were used as an input for LR, and the results from LR are used again with different variable sets in order to enhance the prediction score.

(3) **Genetic Operators:** Operators guarantee the possibility of diversifying populations over the generations and exploring the solution space. We apply the following operations during a GA cycle: Selection, Crossing and Mutation.

(a) **Selection:** The selection consists of choosing the individuals serving to create the next generation, the individuals who will survive. The selection of individuals is carried out most often on the basis of the evaluation function. Several selection operators are used such as the roulette wheel selection [23], Rank in the population [24] or tournament selection [25].

(b) **Crossover:** Crossing is responsible for constructing an individual solution for the problem from the mixture of many other solutions. In crossing, the chromosomes exchange sequences of genes between them. This process is applied to each pair of chromosomes selected with a certain probability of P. The pairs of chromosomes are copied without modification into the next generation with the probability 1 - P. The higher P, the more new individuals appear in the population.

We present the best-known ones among the most used crossover methods: the one-point crossover and the multi-point crossover.

**One-Point Crossover :** It is about randomly choosing a crossing point for each pair of chromosomes and performing a swap of the sets of sequences of this point between the two parents, giving birth to two new offspring as shown in Fig. 5.

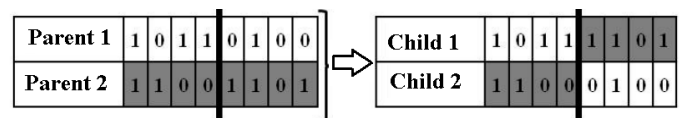


Fig. 5. Single-Point Crossover.

**Multi-Point Crossover :** In this case, several crossing points are selected and the swap is done on the different parts of the sequences surrounded by these points between the genes of the parents as shown in Fig. 6.

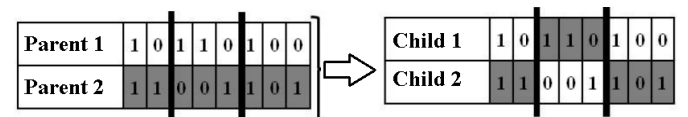


Fig. 6. Multipoint Crossover.

(c) **Mutation :** Mutation is defined as the unexpected change in the value of a gene in a chromosome. Fig. 7 illustrates

an example of a mutation applied to the fourth position of a binary chromosome. The mutation plays the role of noise which prevents the evolution from stopping. It allows the extension of space exploration and guarantees that the global optimum can be reached. This operator, therefore, avoids a convergence towards the local optimum.



Fig. 7. Exemple of Mmutation.

The following section will be dedicated to presenting the results and scores obtained by applying the methods presented in this section on the data obtained in the fourth section .

### V. RESULTS

After performing a hybrid feature selection method in Section 4, we end up with two datasets: a first dataset containing all extracted features (All) and a second one containing only the seven selected features (selected). To assess the relevance of this selection, we will compare the Accuracy and AUC-ROC scores for the two datasets using DL and GA models.

TABLE VIII. MODELS SCORES

Model	Features	Accuracy	AUC ROC
GA	All	0.926	0.898
	Selected	0.933	0.894
DL	All	0.943	0.876
	Selected	0.938	0.887

According to the results obtained in Table VIII, we notice that the scores remained almost the same even after eliminating several features. It is explained by the fact that some eliminated features had no role in predicting the target variable. In other cases, they may introduce noise. The score has been improved for the GA model after eliminating the unnecessary variables.

We also notice that the two algorithms used in this work outperform the basic algorithms used in previous work. It is explained by GA evolutionary and self-correcting character and DL methods.

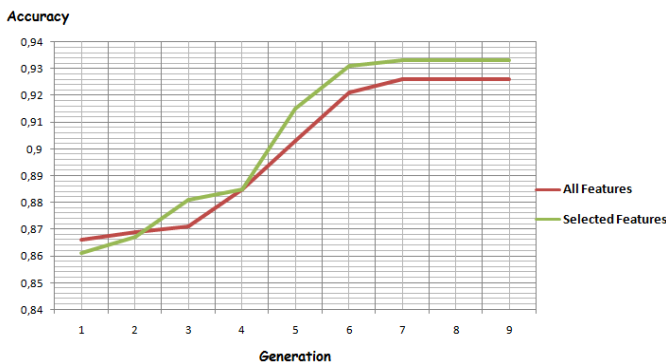


Fig. 8. Accuracy Evolution over Generations.

Fig. 8 provides information about how quickly the GA converges to the optimal solution. After only seven generations, the algorithm found an optimal solution for the problem. This rapid evolution of accuracy over generations depends highly on the value of mutation rate. In practice; it consists of a high mutation rate at the start of the algorithm to allow better solutions for space exploration. Then a decrease in this rate allows the convergence of the algorithm.

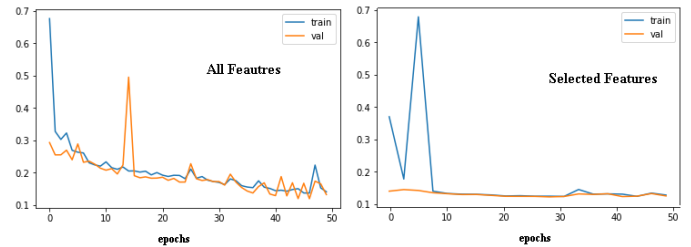


Fig. 9. Loss Evolution over Epochs

Learning curve graphs presented in Fig. 9 are commonly used for the NN model. It plots the variation of loss or accuracy over epochs. The data is divided into two parts training and validation sets. Learning curve graphs are firstly used to examine the model convergence; we expect that the loss decreases and the accuracy increase as the number of epochs increases. We also expect that the model will converge after training for several epochs. Secondly, It is used to diagnose if the model has over-fitted or under-fitted the learning set.

Fig. 9 show that we obtain an accuracy of 90% after 40 epochs for the dataset using all features. The same accuracy is obtained after only 10 epochs for the dataset using selected features, which is explained by the fact that a model containing fewer features will be less complex and require fewer epochs to converge. Fig. 9 show that for both datasets, we can safely stop the training process at 50 epochs without fearing over-fitting or under-fitting.

After proving the relevance of models and the set of features used in this article, and since the temporal aspect is present in our dataset, the next step in this work is to use ARIMA to predict independent variables for future weeks for every enrolment. After obtaining these values, we use them as an observation for the models to predict the value of the target variable Y. With the function predict.proba() in python sickit-learn library, we can find the weekly drop-out percentage.

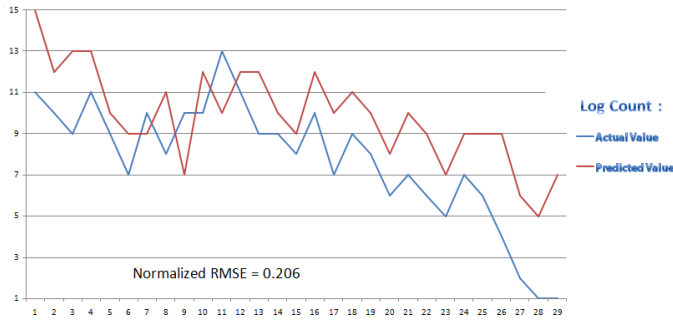


Fig. 10. Log Count Error Prediction

We can judge the accuracy of the ARIMA model predictions through the normalized RMSE (Root Mean Squared Error) value. Fig. 10 shows an example of the error between the actual value of the number of logs per day and the predicted value. The normalized RMSE value is suitable for all variables and indicates the accuracy of the ARIMA model and stationary time series of the variables.

At our university, we extracted student logs from the university MOOC to build a dataset similar to the KDD cup dataset. This dataset has undergone the same process mentioned in this article, starting with feature extraction and feature engineering and ending with models' predictions score. Table IX presents the different model scores for the university dataset.

TABLE IX. MODELS SCORES UNIVERSITY DATA

Model	Features	Accuracy	AUC ROC
GA	All	0.931	0.876
	Selected	0.921	0.887
DL	All	0.888	0.875
	Selected	0.898	0.886

An intelligent system was hosted in the university server based on a single web page using Python and Streamlit framework. The system inputs the enrolment entry (student, course), fetches the corresponding logs and aggregates them by week. In addition, the ARIMA method is used to predict new observations for the following weeks using the previous ones. The system offers the possibility to choose the model used to predict the drop-out rate each week, as shown in Fig. 11 and Fig. 12.

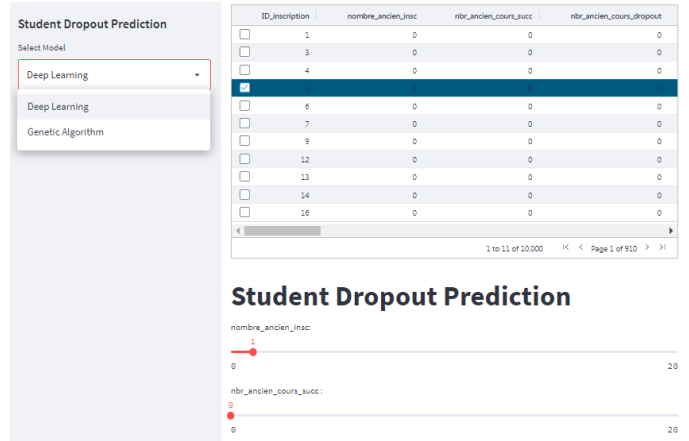


Fig. 11. Web Page Interface.

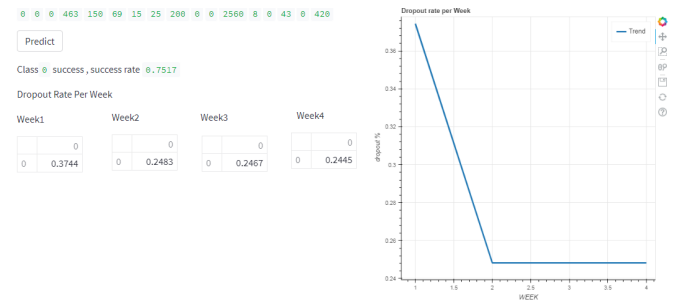


Fig. 12. Results Web Page.

In order to determine student drop-out profiles and patterns, We performed a correlation analysis between extracted features and the student's final performance (completion or drop-out). We found that when the pattern is: Access — Video — Assignments — Discussion is respected, The accuracy of retention is increased visibly. It means that the expected behaviour of the student is to access the course link represented by the event (Access), watch the course content through videos, and then visit the Assignments page. Finally, discuss the ambiguous and misunderstood points on the discussions page. Similarly, the student behaviour on assignments pages, the count of the viewing Video events, Discussion, and wiki page measured weekly, provided an indicator of student engagement and persistence and were an excellent early predictor of the drop-out rate and performance.

When we used K-mean clustering, we found three main clusters for droppers: According to our analysis of weeks 1 and 2, data indicates three dominant clusters.

- **cluster 1** student with little time spent watching videos. This cluster corresponds to students who didn't complete the course videos. This grouping was a strong indicator of drop-out 95% of these students did not complete the course.
- **cluster 2** concerns students who complete watching videos but have a few visits and time spent on the assignments page. In other words, these students have completed most of the course but didn't take quizzes and exercises. This



grouping was also a strong indicator of drop-out: 80% dropped out in the fourth week.

The reasons which explain the behaviour of the students of clusters 1 and 2 can be linked to several causes extracted from related works:

- **causes linked to student:**

- The lack of students motivation and engagement: is considered one of the most influential factors preventing students from completing a MOOC. For instance, [26] surveyed 134 students who had not completed the MOOC courses and found that the majority of students had the intention to complete their study but they were unable to do so due to low motivation and poor feedback.
- The students lack the abilities/skills and prerequisites to follow the course. According to previous studies [27], [28], [29], demonstrating the effect of students' academic skills and abilities and their prior experience on the drop-out rate in MOOCs.

- **causes linked to the course:**

- It can also be linked to the course length and difficulty. According to [30], [31], [32] the complexity or the difficulty of the course content was found in many studies to be associated with students drop-out rates

- **causes linked to instructors:**

- The lack of instructor supervision: The poor feedback provided by the instructors due to the massive number of enrolled students per course has been reported to be an significant predictor of students drop-out in MOOC courses [33] and [34].

- **cluster 3** concerns students who have spent an average time on the videos and assignments page but have few visits to the discussion page. It can be interpreted that those students are not social enough to communicate with others students or ask questions about ambiguous points in courses or exercises.

Isolation and lack of interactivity in MOOCs directly affect students drop-out. A survey [29] about MOOCs drop-out showed from the droppers' comments that they mentioned feeling isolated and unmotivated to continue due to low interaction and communication with students and instructors. They complained that the instructor did not praise or motivate them after the quizzes. They also stated that instructors did not engage learners in discussion or facilitate brainstorming.

After detecting patterns of the student at high risk of drop-out, the following section outlines examples of interventions:

- **Interventions for cluster 1 and 2:** For students lacking the necessary prerequisites to take the course, it is wise to detect them through a survey or quizzes at the beginning of each course. Then send them courses and exercises containing the prerequisites they lack to follow the course. According to the literature, there is a myriad of interventions aiming to increase students motivation and engagement by creating interest in the course topics [35]. Some interventions dealt with demotivation through an email mechanism [36]. Other interventions try to get absent students back into the course and collect their reasons for

leaving [37].

The lack of instructor supervision and the poor instructor feedback is due to the considerable enrolment number per course in MOOCs. The solution here is to focus on students flagged at high risk of drop-out on a particular week according to the drop-out rate given by the system.

Regarding the course content, the course must appear helpful for the students in real life. It should contain a lot of application and practical exercises. The skills learned in the course must apply to real-world problems, particular career goals, or later life roles.

- **Interventions for cluster 3:** Multiple research has found that social connectedness to school is linked to higher rates of student academic success [38], teachers and peers can serve as sources for facilitating this social connection. The intervention proposed for this cluster is a weekly peer-support group meeting that focuses on enhancing students' academic and interpersonal skills combined with daily interactions. It will improve outcomes for students flagged as a potential drop-out. We could form a peer-support group of three to four participants, and the 5th is the student flagged as a drop-out. This methodology is more suitable for blended learning; it has been tested in the university, and the results obtained improved classroom behaviour, increased academic engagement, and positive peer and teacher interactions.

The cost of a students drop-out is very high in terms of wasted time, effort and money. When a student decides to leaves, connection with that student is lost, and generally nothing is done to determine the reasons behind. Institutions can implement this system or similar, to anticipate and reduce the number of drop-out.

Several other drop-out patterns might be detected, and such predefined intervention strategies can be learned from expert teachers or from historical data [12]. The current study is just a first step toward an ultimate automated personalized intervention system.

From an algorithmic perspective, the experiment in this study showed that deep learning is outperforming other baseline algorithms either in prediction accuracy or in generating more accurate drop-out probabilities. Moreover, deep learning showed more robustness against over-fitting.

## VI. CONCLUSION

The excessive drop-out rate in MOOCs encourage to use data mining techniques and ML algorithms in order to predict students at risk of drop-out. In the light of the obtained results we can conclude that for classification problems based on raw activity records, features extraction and data preparation is a necessary step before building models. The hybrid features selection algorithm adopted in this work is effective. One of our main contributions was obtaining competitive prediction results with a minimum number of variables.

According to the result obtained we can also conclude that our proposed models using GA and DL are producing very competitive results in this problem. Models used in this study are outperforming the obtained result using the baseline algorithms in previous works. This study was very useful, and optimises the drop-out prediction in the university MOOC,

because it is not only focusing on early detecting students at risk of drop-out, but it also personalise intervention and seeks for the reasons behind, in order to increase retention rate in the MOOC. As a perspective, the methodology used in this article must be tested on other benchmark datasets in order to assess its relevance.

#### ACKNOWLEDGMENT

We would like to gratefully acknowledge the organizers of KDD Cup 2015 as well as XuetangX for making the datasets available.

#### REFERENCES

- [1] D. Lizcano, J. A. Lara, B. White, and S. Aljawarneh, "Blockchain-based approach to create a model of trust in open and ubiquitous higher education," *Journal of Computing in Higher Education*, vol. 32, no. 1, pp. 109–134, 2020.
- [2] D. Shah, "The second year of the mooc: A review of mooc stats and trends in 2020," *The Report by class central [Electronic resource]*. URL: <https://www.classcentral.com/report/the-second-year-of-the-mooc/> (accessed: 09.11. 2021), 2020.
- [3] F. Dalipi, A. S. Imran, and Z. Kastrati, "Mooc dropout prediction using machine learning techniques: Review and research challenges," in *2018 IEEE global engineering education conference (EDUCON)*. IEEE, 2018, pp. 1007–1014.
- [4] C. Bossu and T. Heck, "Engaging with open science in learning and teaching," *Education for Information*, vol. 36, no. 3, pp. 211–225, 2020.
- [5] R. Yu, H. Lee, and R. F. Kizilcec, "Should college dropout prediction models include protected attributes?" in *Proceedings of the eighth ACM conference on learning@ scale*, 2021, pp. 91–100.
- [6] A. F. Wise, S. Knight, and S. B. Shum, "Collaborative learning analytics," in *International handbook of computer-supported collaborative learning*. Springer, 2021, pp. 425–443.
- [7] C. F. de Oliveira, S. R. Sobral, M. J. Ferreira, and F. Moreira, "How does learning analytics contribute to prevent students' dropout in higher education: A systematic literature review," *Big Data and Cognitive Computing*, vol. 5, no. 4, p. 64, 2021.
- [8] J. Figueroa-Cañas and T. Sancho-Vinuesa, "Changing the recent past to reduce ongoing dropout: an early learning analytics intervention for an online statistics course," *Open Learning: The Journal of Open, Distance and e-Learning*, pp. 1–18, 2021.
- [9] P. M. Moreno-Marcos, C. Alario-Hoyos, P. J. Muñoz-Merino, and C. D. Kloos, "Prediction in moocs: A review and future research directions," *IEEE transactions on Learning Technologies*, vol. 12, no. 3, pp. 384–401, 2018.
- [10] N. Gitinabard, F. Khoshnevisan, C. F. Lynch, and E. Y. Wang, "Your actions or your associates? predicting certification and dropout in moocs with behavioral and social features," *arXiv preprint arXiv:1809.00052*, 2018.
- [11] J. Berens, K. Schneider, S. Görtz, S. Oster, and J. Burghoff, "Early detection of students at risk—predicting student dropouts using administrative student data and machine learning methods," *Available at SSRN 3275433*, 2018.
- [12] W. Xing and D. Du, "Dropout prediction in moocs: Using deep learning for personalized intervention," *Journal of Educational Computing Research*, vol. 57, no. 3, pp. 547–570, 2019.
- [13] Z. Alj and M. O. C. Malki, "Predicting students drop-out in mooc from learning behavior using machine learning," *Journal of Uncertain Systems*, p. 2250011, 2022.
- [14] R. Munoz, R. Olivares, C. Taramasco, R. Villarroel, R. Soto, T. S. Barcelos, E. Merino, and M. F. Alonso-Sánchez, "Using black hole algorithm to improve eeg-based emotion recognition," *Computational intelligence and neuroscience*, vol. 2018, 2018.
- [15] R. Olivares, R. Munoz, R. Soto, B. Crawford, D. Cárdenas, A. Ponce, and C. Taramasco, "An optimized brain-based algorithm for classifying parkinson's disease," *Applied Sciences*, vol. 10, no. 5, p. 1827, 2020.
- [16] Y. Xue, H. Zhu, J. Liang, and A. Stowik, "Adaptive crossover operator based multi-objective binary genetic algorithm for feature selection in classification," *Knowledge-Based Systems*, vol. 227, p. 107218, 2021.
- [17] M. LLC. (1999) MS Windows NT kernel description. [Online]. Available: [https://scikit-learn.org/stable/modules/generated/sklearn.feature\\_selection.RFE.html](https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.RFE.html)
- [18] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.
- [19] H. He, Y. Bai, E. A. Garcia, and S. Li, "Adasyn: Adaptive synthetic sampling approach for imbalanced learning," in *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, 2008, pp. 1322–1328.
- [20] X. Geng, T.-Y. Liu, T. Qin, and H. Li, "Feature selection for ranking," in *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, 2007, pp. 407–414.
- [21] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artificial intelligence*, vol. 97, no. 1-2, pp. 273–324, 1997.
- [22] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of machine learning research*, vol. 3, no. Mar, pp. 1157–1182, 2003.
- [23] D. E. Goldberg and K. Deb, "A comparative analysis of selection schemes used in genetic algorithms," in *Foundations of genetic algorithms*. Elsevier, 1991, vol. 1, pp. 69–93.
- [24] G. Syswerda, "A study of reproduction in generational and steady-state genetic algorithms," in *Foundations of genetic algorithms*. Elsevier, 1991, vol. 1, pp. 94–101.
- [25] B. L. Miller, D. E. Goldberg *et al.*, "Genetic algorithms, tournament selection, and the effects of noise," *Complex systems*, vol. 9, no. 3, pp. 193–212, 1995.
- [26] C. Gütl, R. H. Rizzardini, V. Chang, and M. Morales, "Attrition in mooc: Lessons learned from drop-out students," in *International workshop on learning technology for education in cloud*. Springer, 2014, pp. 37–48.
- [27] H. B. Shapiro, C. H. Lee, N. E. W. Roth, K. Li, M. Çetinkaya-Rundel, and D. A. Canelas, "Understanding the massive open online course (mooc) student experience: An examination of attitudes, motivations, and barriers," *Computers & Education*, vol. 110, pp. 35–50, 2017.
- [28] M. Yamba-Yugsi and S. Luján-Mora, "Cursos mooc: factores que disminuyen el abandono en los participantes," *Enfoque UTE*, vol. 8, pp. 1–15, 2017.
- [29] K. S. Hone and G. R. El Said, "Exploring the factors affecting mooc retention: A survey study," *Computers & Education*, vol. 98, pp. 157–168, 2016.
- [30] G. R. El Said, "Understanding how learners use massive open online courses and why they drop out: Thematic analysis of an interview study in a developing country," *Journal of Educational Computing Research*, vol. 55, no. 5, pp. 724–752, 2017.
- [31] T. Eriksson, T. Adawi, and C. Stöhr, "'time is the bottleneck': a qualitative study exploring why learners drop out of moocs," *Journal of Computing in Higher Education*, vol. 29, no. 1, pp. 133–146, 2017.
- [32] S. Zheng, M. B. Rosson, P. C. Shih, and J. M. Carroll, "Understanding student motivation, behaviors and perceptions in moocs," in *Proceedings of the 18th ACM conference on computer supported cooperative work & social computing*, 2015, pp. 1882–1895.
- [33] S. Halawa, D. Greene, and J. Mitchell, "Dropout prediction in moocs using learner activity features," *Proceedings of the second European MOOC stakeholder summit*, vol. 37, no. 1, pp. 58–65, 2014.
- [34] D. F. Onah, J. Sinclair, and R. Boyatt, "Dropout rates of massive open online courses: behavioural patterns," *EDULEARN14 proceedings*, vol. 1, pp. 5825–5834, 2014.
- [35] T. NeCamp, J. Gardner, and C. Brooks, "Beyond a/b testing: sequential randomization for developing interventions in scaled digital learning environments," in *Proceedings of the 9th International Conference on learning analytics & knowledge*, 2019, pp. 539–548.
- [36] I. Borrella, S. Caballero-Caballero, and E. Ponce-Cueto, "Predict and intervene: Addressing the dropout problem in a mooc-based program," in *Proceedings of the Sixth (2019) ACM Conference on Learning@ Scale*, 2019, pp. 1–9.
- [37] J. Whitehill, J. Williams, G. Lopez, C. Coleman, and J. Reich, "Beyond prediction: First steps toward automatic intervention in mooc student stopout," *Available at SSRN 2611750*, 2015.

- [38] L. Bond, H. Butler, L. Thomas, J. Carlin, S. Glover, G. Bowes, and G. Patton, "Social and school connectedness in early secondary school as predictors of late teenage substance use, mental health, and academic outcomes," *Journal of adolescent health*, vol. 40, no. 4, pp. 357–e9, 2007.

# An Intelligent Decision Support Ensemble Voting Model for Coronary Artery Disease Prediction in Smart Healthcare Monitoring Environments

ANAS MAACH<sup>1</sup>

LASTIMI, High School of Technology  
Mohammed V University in Rabat  
Sale, Morocco

JAMILA ELALAMI<sup>2</sup>

LASTIMI, High School of Technology  
Mohammed V University in Rabat  
Sale, Morocco

NOUREDDINE ELALAMI<sup>3</sup>

LASTIMI, Mohammedia School of Engineers  
Mohammed V University in Rabat  
Rabat, Morocco

EL HOUSSINE EL MAZOUZI<sup>4</sup>

CISIEV, Faculty of Science and Technology in Marrakech  
Cadi Ayyad University  
Marrakech, Morocco

**Abstract**—Coronary Artery Disease (CAD) is one of the most common cardiac diseases worldwide and causes disability and economic burden. It is the world's leading and most serious cause of mortality, with approximately 80% of deaths reported in low- and middle-income countries. The preferred and most precise diagnostic tool for CAD is angiography, but it is invasive, expensive, and technically demanding. However, the research community is increasingly interested in the computer-aided diagnosis of CAD via the utilization of machine learning (ML) methods. The purpose of this work is to present an e-diagnosis tool based on ML algorithms that can be used in a smart healthcare monitoring system. We applied the most accurate machine learning methods that have shown superior results in the literature to different medical datasets such as RandomForest, XGBoost, MultilayerPerceptron, J48, AdaBoost, NaiveBayes, LogitBoost, KNN. Every single classifier can be efficient on a different dataset. Thus, an ensemble model using majority voting was designed to take advantage of the well-performed single classifiers, Ensemble learning aims to combine the forecasts of multiple individual classifiers to achieve higher performance than individual classifiers in terms of precision, specificity, sensitivity, and accuracy; furthermore, we have benchmarked our proposed model with the most efficient and well-known ensemble models, such as Bagging, Stacking methods based on the cross-validation technique, The experimental results confirm that the ensemble majority voting approach based on the top three classifiers: MultilayerPerceptron, RandomForest, and AdaBoost, achieves the highest accuracy of 88,12% and outperforms all other classifiers. This study demonstrates that the majority voting ensemble approach proposed above is the most accurate machine learning classification approach for the prediction and detection of coronary artery disease.

**Keywords**—Machine learning; smart healthcare; coronary artery disease

## I. INTRODUCTION

No cure exists for Coronary Artery Disease (CAD), as a combination of environmental and inherited factors is thought to be associated with several risk factors, including a family history of heart disease, age, overweight, inactivity, poor diet, and tobacco usage. The diagnosis of coronary artery disease is very challenging for the General Physician (GP). When a

patient experiences chest pain, he consults the GP. The chest pain is the main reason for consultation in approximately 4% of cases and in only 15% of cases [1], Coronary Artery Disease (CAD) ultimately will be diagnosed as the reason for the symptoms, The difficulty for the GP is to identify CAD on the basis of symptoms, age, and gender. Distinguishing a life-threatening disease from a non-life-threatening disease is crucial for the effective prevention and management of the disease, but it can be difficult, especially in cases of atypical blood pressure or non-specific chest complaints [1], [2]. Currently, approximately 105,000 people 53% of whom are women are being recommended to a cardiologist, each year in the Netherlands. In fact, only 5% of males and 1% of females have coronary disease needing invasive therapy. Therefore, a clinical need exists in the population suffering from chest pain for optimizing the diagnosis and orientation to the cardiologist [3]. Thus, the development of an accurate diagnostic tool based on ML algorithms would assist GPs in identifying the likelihood of coronary artery disease and in guiding the management of patients. In addition, early diagnosis of chronic diseases saves the expense of medical care and reduces the likelihood of more complex health problems, especially considering the lack of doctors in underserved areas and developing countries. In this case, the association of Wireless Body Area Networks (WBANs) and machine learning methods should be used to assist practitioners in the early diagnosis and identification of CAD by offering predictive models for better and faster decision support. Nevertheless, it is worth noting that machine learning tends to be looked upon with suspicion by some due to what can be termed a “black box” [4]: being unable to reveal its inner decision-making mechanism. However, this inability to explain its inner decision-making tends to lead to skepticism among consumers and slow adoption by end-users in the domain of health care. It is crucial to build trust, especially in healthcare, where errors can be fatal, to be able to convey both the underlying reasoning and the process required to obtain a machine learning prediction. This paper aims to develop a CAD detection, classification, and prediction tool that can be integrated into a smart healthcare system. Through the use of ensemble machine learning approaches combining the best

classifiers Multilayer Perceptron, Adaboost and RandomForest, such a system would be capable of predicting whether a person is likely to have CAD on the basis of various relevant indicators, supplying physicians with an advance diagnostic assessment. The classification models were evaluated using various metrics, namely, F-measure, accuracy, recall, precision, sensitivity, and the ROC curve (receiver operating characteristic curve) to select the most efficient classifier. Several relevant features that can potentially be utilized to predict coronary artery disease have been taken from the best classifier scheme. The Z-Alizadeh Sani dataset [5] is used for the purposes of this work. The Z-Alizadeh Sani dataset comprises 303 records of patients, with 55 features. All the features can be regarded as CAD indicators, as stated in the literature [5]. Features are categorized into four categories: demographics, symptoms, laboratory, and ECG features. Accordingly, every patient may be classified in two different possible classes: CAD or normal. The patient is classified as having CAD if his or her narrowing diameter is greater than 50% and if not, he is normal.

#### A. An Overview of the Smart Healthcare Monitoring System

A Smart Healthcare Monitoring system based on Wireless Body Area Networks (WBANs) is organized into a three-layer telemedicine system as shown in Fig. 1. This system consists of a network of wireless sensors that continuously track the health parameters of patients [6], [7]. Furthermore, this healthcare monitoring system is interconnected to the high-level biomedical server via an internet-based network system.

1) *Tier 1*: is also called the WBAN level, in which, each patient under healthcare monitoring is connected to several small Body Sensor Networks (BSNs), These inhomogeneous sensors are placed either in the body or in wearable devices As shown in Fig. 1, the BSN detects different physiological body parameters. These include electroencephalogram (EEG), pulse rate, electromyography (EMG), blood pressure, electrocardiogram (ECG), and so forth [6], [8], [7]. To communicate within the WBAN level, those BSNs are using radio waves to communicate with themselves and with the coordinator, A sink node is operating as a hub for all the BSNs .

2) *Tier 2* : The second level is implemented in a PC/laptop, mobile phone, or PDA. The data collected from the BSNs in various formats, including graphics, digital, audio and so forth [6], [8], are transferred to a healthcare server. It employs several technologies, such as 4G/5G or WiFi, to communicate with a remotely located medical server.

3) *Tier 3*: It consists of a large network of various devices, services, healthcare practitioners, and healthcare services providers that are interconnected. This layer delivers many services to potentially thousands of clients through the use of healthcare systems as a centralized point of contact. These medical servers store the health data of patients and deliver various additional services to these patients and other related stakeholders [6], [7]. The tasks of the health server involve authentication of patients, acceptance, and submission of their medical data, and formatting and analysis of the data to identify the severity of health problems. If the analyzed data shows that the patient's potential medical condition is of a life-threatening type, the health server alerts emergency caregivers. Patients and their doctors can access the analyzed data at their

location via the Internet. Patient data is reviewed by doctors to assess whether it is in accordance with the desired healthy ranges (e.g. pulse pressure, heart rate, etc.) and whether the given or prescribed medical treatment is working. The rest of the article is organized in the following sections: Section 2 provides a review of the state-of-the-art literature on coronary heart disease and heart disease research, while Section 3 presents the proposed methodology, Section 4 reports the results of the experiments and the discussion, The limitations of this article and future work are described in Section 5, and Section 6 provides a conclusion to the article.

## II. RELATED WORKS

Up to now, various research studies have been carried out on the early diagnosis of coronary artery disease and heart disease. They have utilized various machine learning prediction approaches and achieved remarkable performance. This section provides an extensive literature review of research studies in the field of heart disease diagnosis supported by machine learning techniques: In [9] Yadav et al. presented a novel method for ensemble machine learning utilizing Pearson correlation and chi-square feature selection-based algorithms for the correlation strength of heart disease attributes and the Random Forest ensemble method for the diagnosis of heart disease. The authors performed experiments with their proposed system on the CHDD dataset, and they were able to achieve the best performance considering many evaluation metrics Correctly Classified Instances, Mean absolute error, Incorrectly Classified Instances, Kappa statistic, Root relative squared error, Relative absolute error, and root mean squared error, the Random Forest ensemble method outperforms various machine learning techniques RF, AdaBoostM1, Gradient Boosting. In [10] Li et al. proposed a high-performance and intelligent approach for detecting cardiac diseases, and the model is based on a feature selection method (FCMIM) with a support vector machine classifier (SVM). They conducted experiments with their proposed method on the CHDD dataset and were able to achieve the best performance considering many evaluation parameters: accuracy, MCC specificity, processing time, and sensitivity against various machine learning techniques SVM, LR, ANN, kNN, NB, and DT. In [11] Javeed et al. introduced a new heart failure prediction diagnostic method using a random search algorithm (RSA), which is applied for feature selection, and a random forest model to perform classification and prediction. They carried out experiments with their proposed system on the CHDD dataset and were able to achieve the best accuracy, sensitivity, specificity, and MCC. The proposed method outperforms various machine learning techniques, including the random tree model, the Adaboost model, SVM with a linear kernel function, the additional tree ensemble model, and the support vector machine (RBF kernel). In [12] Saxena et al. presented an innovative automated system that combines Generalized Discriminant Analysis (GDA) as an effective feature reduction algorithm together with a Radial Basis Function (RBF) kernel and Online Sequential Extreme Learning Machine (OSELM) based on a Sigmoid activation function, Hardlim, RBF and Sine as a binary classifier for the detection of congestive heart failure (CHF) and coronary artery disease (CAD). They performed experiments with their proposed system on the NSR-CAD, NSR-CHF, and CAD-CHF datasets and were able to obtain the best results in terms of

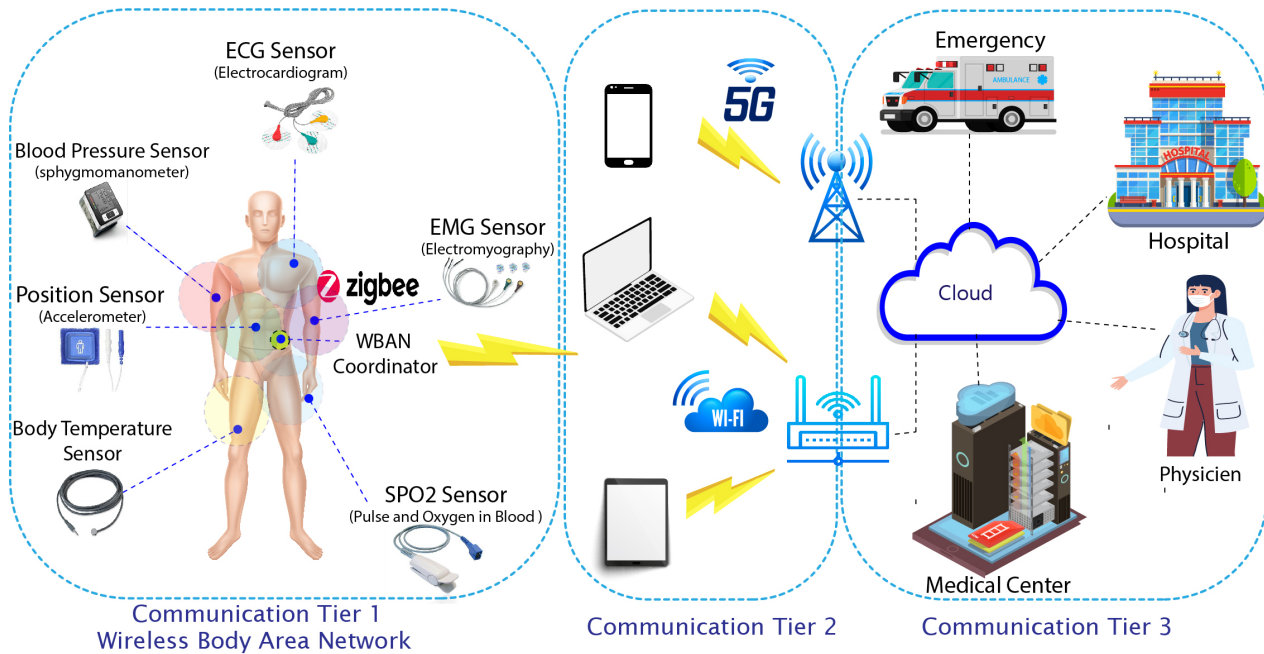


Fig. 1. General Architecture of a Smart Healthcare Monitoring System.

accuracy, sensitivity, specificity, mean  $\pm$  SD and p-value. The proposed method outperforms the different machine learning techniques GDA-Kernel Function (Gaussian, Polynomial, and RBF) and OS-ELM (RBF activation, Sigmoid, Hardlim, and Sine). In [13] Dwivedi et al. proposed an approach for accurately diagnosing heart disease using the logistic regression method. They conducted experiments with their proposed technique on the StatLog heart disease dataset and were able to achieve the best performance considering many evaluation parameters: classification accuracy, precision, F1-measure, false positive rate (FPR), sensitivity, specificity, negative predictive value (NPV), misclassification rate (MRR), compared to five different data mining techniques: ANN, SVM, kNN, CT, and NB. In [14] Gupta et al. presented an intelligent decision-support model that can help medical experts in predicting heart disease through an advanced ensemble classifier. They conducted experiments with their proposed system on the CHDD dataset and were able to achieve the best performance considering many evaluation parameters: classification accuracy, specificity, F-measure, recall accuracy, MAE, ROC, and RMSE against various machine learning techniques: MLP, NB, J48, RF, SVM, AB, boosted tree, and binary discriminant. In [15] Verma et al. presented a new hybrid method for the diagnosis of CHD, including the identification of risk factors using correlation-based feature subset selection (CFS) with particle optimization search (PSO) and K-means clustering approaches. They conducted experiments with their proposed system on the CHDD dataset and they were able to obtain the best model performance against five different machine learning techniques: MLP, MLG, FURI, and DT (C4.5). In [16] Miao et al. proposed an improved ensemble machine learning scheme using an adaptive boosting algorithm for accurately diagnosing Coronary Artery Disease (CAD). They have performed experiments with their proposed method on

CHDD, HHDD, LBMC, and SUH datasets and they were able to achieve the best performance considering many evaluation parameters: accuracy, ROC, classification error, precision, sensitivity (or recall), F-score, K-S measure, specificity, and AUC, and to outperform other machine learning techniques. In [17] Long et al. proposed a heart disease diagnostic system using rough set-based feature reduction and type 2 fuzzy logic systems (IT2FLS) for early stage heart disease detection, in which the authors implemented the BPSORS-AR Binary Particle Swarm Optimization and the rough set-based feature selection technique. They conducted experiments with their proposed model on the heart disease dataset and the SPECTF dataset, and they were able to achieve the best performance compared to the different data mining techniques, NB, SVM, and ANN. In [18] Nilashi et al. proposed a new methodology for heart disease diagnosis using machine learning algorithms. Such a model was built with unsupervised and supervised machine learning methods. using the implementation of Fuzzy Logic and the Support Vector Machine (SVM)-based ensemble model, and Principal Component Analysis (PCA) was employed with two processes for imputation. Both imputation methods were essentially used for missing value imputation. In addition, they have implemented the Augmented FSVM and Augmented PCA for augmented learning of the data. This was done to reduce the computational time. This was associated with the prediction of the disease. According to the results, it was deduced that the ensemble model showed high accuracy in classifying heart disease and also decreased the computational time required for disease diagnosis. From this brief state of the art, it can be deduced that there is a great interest in the scientific community for the prediction and detection of heart diseases, but in reality, machine learning tools are not yet widely applied in diagnostic systems for heart diseases, especially in developing countries, where the mortality rate



from heart disease is very high. This is because many proposed schemes are too complex to be implemented in a smart healthcare monitoring system for heart disease; hence, there is still space for improvement. In this work, a new approach based on a majority voting ensemble model that combines the prediction of three classifiers (Adaboost, Multilayer Perceptron, Random Forest) is proposed. Unlike other approaches, this approach is simple to implement and gives excellent results in the detection and prediction of coronary artery disease.

### III. PROPOSED METHODOLOGY

In Section 3, we provide a description of the proposed methodology and also explain that the proposed approach is a process defined by the following steps, as shown in Fig. 3.

#### A. Dataset Pre-Processing

1) *Dataset Description:* A variety of experiments were performed utilizing the Z-Alizadeh Sani dataset. Originally, this dataset was provided by the Shaheed Rajaei Cardiovascular Medical and Research Center. It was constructed from the records of 303 random visitors, The reason for choosing this dataset is that it includes clinical and non-clinical features that can be gathered remotely via WBANS, such as ECG features (EF-TTE, RWMA, Q-wave, and T-inversion), in contrast to other datasets and which have a significant impact on the prediction of coronary artery disease. The predictor variables are Age, Diabetes Milletus(DM), Hypertension(HTN), Blood Pressure(BP), Typical Chest Pain, Atypical, Nonanginal, T-inversion, Fasting Blood Sugar(FBS), Erythrocyte Sed rate(ESR), Potassium(K), Ejection Fraction(EF-TTE), Regional Abnormality(Region RWMA), as shown in Table I The dataset consists of 303 instances, divided into 216 CAD instances and 87 healthy instances The target variable identifies whether a person has CAD, represented by 1, or not, represented by 0 The description of the parameters of each attribute in the dataset, including the mean, the median, the maximum and minimum, and the value of the standard deviation, is presented in Table II.

2) *Data Cleaning :* Data cleaning is the following phase of the machine learning process. It is regarded as a key step in the workflow process of our approach because it either builds the model or breaks it. Different aspects of data cleaning need to be considered:

- Noise, Duplicates, Invalid or missing data
- Normalization
- Filter unwanted outliers
- Deal with imbalanced datasets.

a) *Dealing with Outliers :* It is essential to identify faulty measurements (outliers) that are divergent from other measurements and to detect sensor faults in emergency situations in order to minimize false alarms. Anomalous measurements should be excluded in order to minimize unnecessary false alarms and interventions by health professionals. As shown in Fig. 7 and Fig. 6, there are abnormal measurements (outliers) in the Fasting Blood Sugar (FBS), Erythrocyte Sed Rate (ESR), and Potassium (K), or extreme values in the Fasting Blood Sugar (FBS). Therefore, to extract the outliers and

extreme values, we have proceeded to apply the interquartile range filter. The Fig. 2 shows the number of outliers and extreme values found in the dataset. In order to solve the problem of outlier values and extreme values, and since there are not many instances in the dataset (only 303), we proceeded to standardize the features.

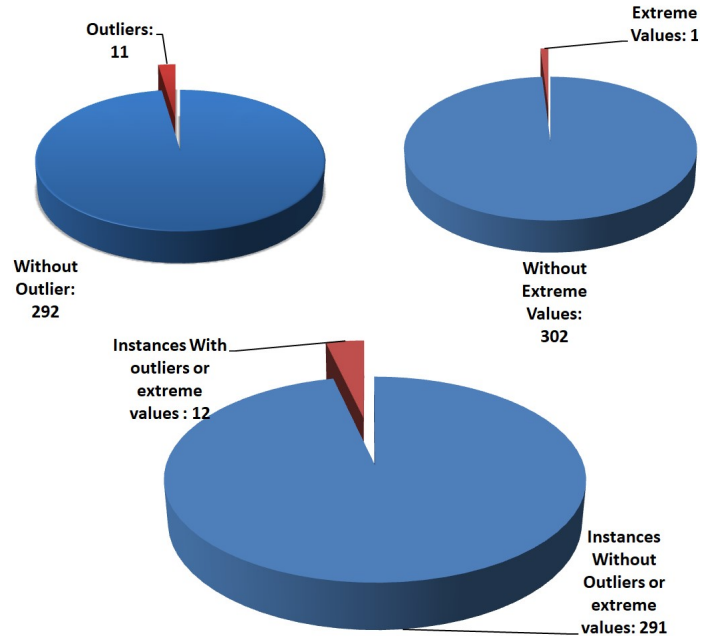


Fig. 2. Outliers and Extremes Values Percentages in the Data.

b) *Dealing with Imbalanced Dataset :* The unbalanced class instances in the health dataset are a critical issue. In fact, the dataset that was employed in our classification experiment was unbalanced since the instances of the first class exceeded the instances of the second class by a significant ratio, which means that the instances are not adequately distributed among the different classes. Therefore, the results of the classification from unbalanced class data produce a biased outcome in favor of the dominant class as shown in Fig. 4. In order to balance the unbalanced dataset, there exist two main techniques, namely, oversampling and undersampling. In the Z-Alizadeh Sani dataset used for this work, positive instances exceeded negative ones, which were solved using the SMOTE method as shown in Fig. 5, Out of the variety of oversampling techniques that exist, SMOTE has demonstrated tremendous potential [19] and is thus widely used by scientists in the medical research community. SMOTE is a technique of oversampling proposed to prevent the issue of class imbalance in the dataset. It improves the classifier's performance and joins the lesser class points to the line segments with the unreal points positioned on these lines. With SMOTE, newly created instances are generated by synthetically resampling the minor class data points, as has been performed in the conventional oversampling method [20], [21] This varies from the conventional approach by the fact that it is carried out in the space of features rather than data space, by regard to the instance of the smaller class at its nearest vector [20], [21] The newly created synthetic data parameters may be generated by applying two distinct approaches, One approach employs the ratio of oversampling, whereas another approach employs the k-nearest neighbors.

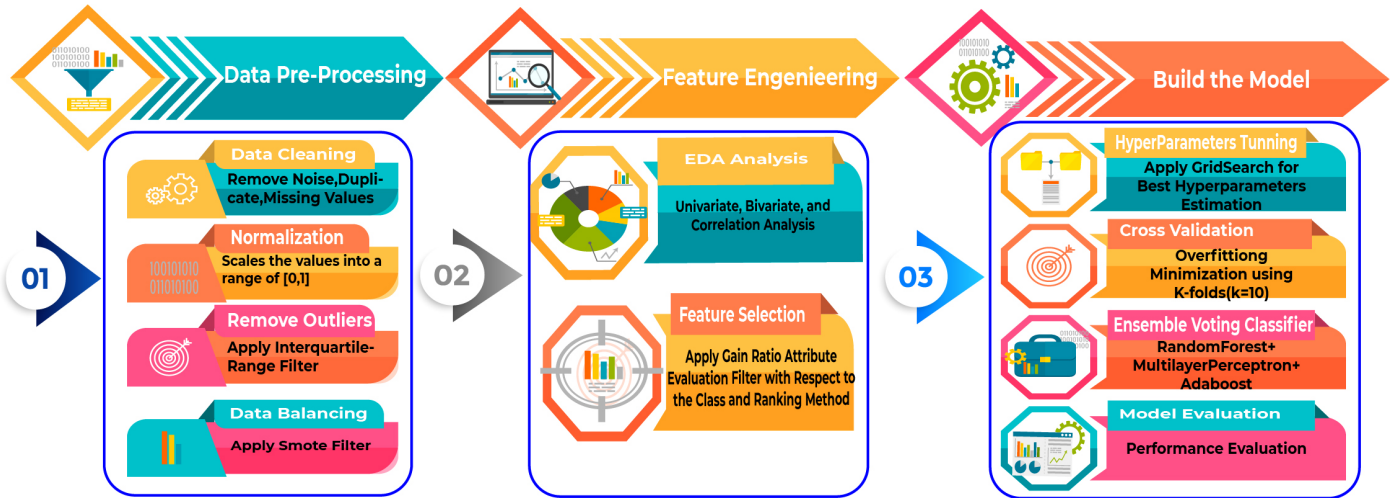


Fig. 3. The Proposed Approach.

This means that SMOTE generates the synthetic data points for the minority class [19] in order to shift the bias of the classifier’s learning from the dominant class to the minor class.

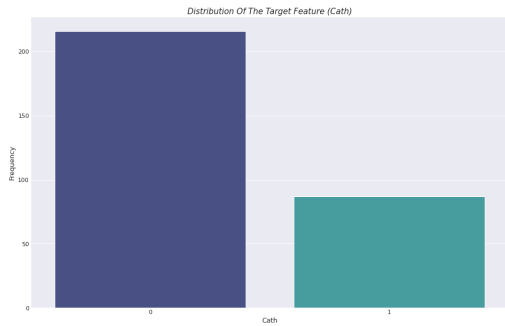


Fig. 4. Distribution of Classes before Applying the Smote Technique.

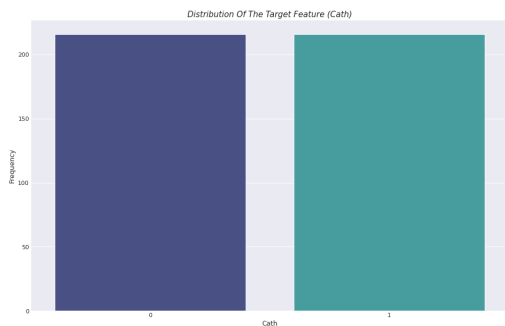


Fig. 5. Distribution of Classes after Applying the Smote Technique.

3) *Features Selection*: In this paper, we employ the Gain Ratio method [22] to identify the most pertinent and useful features. The gain ratio method allows us to check the closeness of features by different methods. The gain ratio provides one of these techniques. It identifies the pertinence of every feature and selects the attributes that have the maximum gain ratio with regard to the likelihood of each feature value. The chosen test must acquire a large gain of information, which

should be inclusive of or larger than the average of the gains of the assessed tests, with the aim of penalizing the spread of the nodes, and must be large when the data is uniformly distributed as well as small if the data belongs to a single branch. Each attribute’s Gain Ratio is computed according to the formula:

$$GainRatio(Attr) = \frac{IG(Attr)}{H(attr)} \quad (1)$$

where

$$H(attr) = \sum -P(val_i) \log_2 P(val_i) \quad (2)$$

and  $P(val_i)$  is defined as the probability of having the value  $val_i$  as a factor of  $t$  global values for a given attribute  $i$

The dataset used contained 55 features; we applied the gain ratio algorithm with various thresholds regarding the number of most relevant attributes that should be utilized in these experiments, and in fact, by using the 12 features, we found the greatest accuracies.

### B. Exploratory Data Analysis

The following section presents a statistical overview of the CAD dataset, described in the Table II. Pair plots provide a simple mechanism to examine how two attributes correlate with each other. Every variable from the dataset is presented in a correlation matrix, which can be immediately visualized. It also provides an effective way to determine the appropriate classification method that should be conducted, Fig. 10 also illustrates the feature distribution in the Coronary Artery Disease dataset, providing a useful representation of the distribution of attributes in the dataset. Fig. 10 represents the plot of all attributes in the dataset (12 attributes). One can observe that three of the attributes, specifically age, K, and BP, are normally distributed. In addition, the dominating value of Tinversion, DM, Atypical, Nonanginal, and RegionRWMA is 0, whereas for the attributes HTN and TypicalChestPain, the

most frequently occurring values are 1 and the least occurring value is 3. Furthermore, Fig. 10 depicts that there were six categorical and six numerical attributes.

1) *Analysis of the Correlation Heatmap*: The Correlation Heatmap is presented in Fig. 8, it is defined as a graphical plot of a cross-correlation matrix that reveals the interrelationship of various attributes. Within the -1 to 1 range, the coefficient of correlation can be given any value. If there is a direct linear relationship between two variables, this relationship is statistically termed a correlation. One can also describe this as a correlation measurement involving two variables. The aim in this case is to identify a correlation between multiple variables and to arrange the results. In this context, a matricial structure of data has been used to store the information. The Fig. 8 shows the correlation feature by feature. In the Fig. 8 we can see various elements of information. First of all, the three features that present the strongest dependency between class and feature are Typical ChestPain, Atypical, Age, RegionRWMA, HTN, Nonanginal and EF-TTE with corresponding correlations of -0.54, 0.42, -0.36, -0.32, -0.29, 0.29, 0.27, and 0.23, respectively. The next fact points out the correlation between two features in HTN-BP, DM-FBS, Atypical-Typical Chest Pain, and RegionRWMA-EF-TTE with corresponding correlations of 0.57, 0.68, -0.72, -0.45, respectively, whereas FBS, K, ESR, BP, and DM have the weakest correlation with the class.

2) *Analysis of The Scatter Plot Matrix*: The scatter plot matrix shown in Fig. 9 is useful for finding pairwise relationships of features. From this, we can deduce the relationships between the features in advance: The more scattered the points, the weaker the relationship, and the more clustered they are, the stronger the relationship. Referring to the scatter plot matrix as shown in Fig. 9 we deduce that there is a relationship between the selected features, such as between DM and FBS, between BP and HTN, and Age.

### C. Methods

In this paper 10, high-level classifiers that showed superior performance in detecting cardiac diseases were selected based on the literature and two ensembles of voting classifiers that we designed by combining a set of three high-level classifiers, the novelty is that this combination, to our knowledge, has never been done before in the literature. The Ensemble voting classifiers are based on the majority voting method for predicting coronary artery disease, the parameters of the Random forest Multilayer Perceptron and Adaboost classifiers have been optimized using the Grid search and CVParameterSelection hyperparameter techniques and eventually, 10 folds cross-validation technique have been utilized to validate the models. The description of the three classifiers that compose our model is presented below:

1) *Adaboost Classifier*: Using the ADAboost classifier is a well-known boosting method. This classifier aids in the consolidation of several weak classifiers into a single effective classifier. Initially, a classifier is fitted on the initial dataset, and then repeated duplicates from the classifier are fitted to the similar dataset, with the weights of erroneously categorized instances modified such that later classifiers concentrate more on challenging situations.

2) *Random Forest (RF)*: Researchers are paying more and more attention to Random Forest, it is an advanced machine learning scheme that demonstrates the overall ensemble learning abilities and ease of use. Both regression and the creation of random subsets require the RD approach. Classification is the principal application of the concept of “bagging” which boosts accuracy rates by mixing learning models, numerous decision trees, of which each is employed in the RF method, make up an ensemble classifier. Since every decision tree is built separately, subsequent trees are intended to be independent of preceding trees [23], each tree in the forest is created to depend on a random vector’s values selected separately using a bootstrapped data set sample, and all the forest trees use the same distribution. In the RF-produced model, random sampling with substitutions is implemented [24]. A random subgroup of the entire set of predictors is used to create the best classifier for each node [25]. The fact that RF uses more computing resources—such as storage spaces—than other algorithms is one of the key shortcomings [26] it addresses, but because of its outstanding prediction accuracy, overfitting avoidance, and scalability it is favored by many researchers.

3) *Multi-Layer Perceptron*: Instead of learning by observation, supervised learning procedures use “learning by example”. To build a learning model, a trained data set has been produced. The learning model is used to test the current input, and predictions, are made. The MLP approach allows for the training of a back propagation-based multilayer feed-forward neural network, which calculates the associated network weights based on the intended outcomes and training patterns. MLP belongs to the class of supervised neural networks that iteratively learn a set of weights for categorical variable prediction [27]. An MLP network’s components are represented by the layers in Fig. 11 input and output layers and several hidden layers [28]. The three parameters of the MLP network may be altered depending on the kind and amount of data. The best prediction should be found by optimizing the parameter’s momentum, learning rate and the number of hidden layers. The learning rate is a measure of how quickly the network is being trained. In other words, when learning rates rise, networks train more quickly but at the risk of creating networks that are unstable. By balancing the network [29], the momentum avoids potential issues that might arise from choosing a fast learning rate that renders instability on the network [30]. Various objective functions and characteristics of input are represented by including the hidden layers.

## IV. EXPERIMENT AND RESULTS

### A. Method of Validation of the Models

This research paper utilized the cross-validation technique with ten folds and four performance evaluation measures. More detail is provided in the subsections below:

1) *Cross-Validation (CV)*: In the present study, a 10-fold cross-validation method [31] is employed to validate the classification model. Aiming to minimize the bias related to selecting random sets from the training data samples while making a comparison of the predictive accuracies of at least two different methods, a k-fold crossvalidation technique was used. In the k-fold cross-validation technique the training dataset S is partitioned randomly into k mutual subsets folds

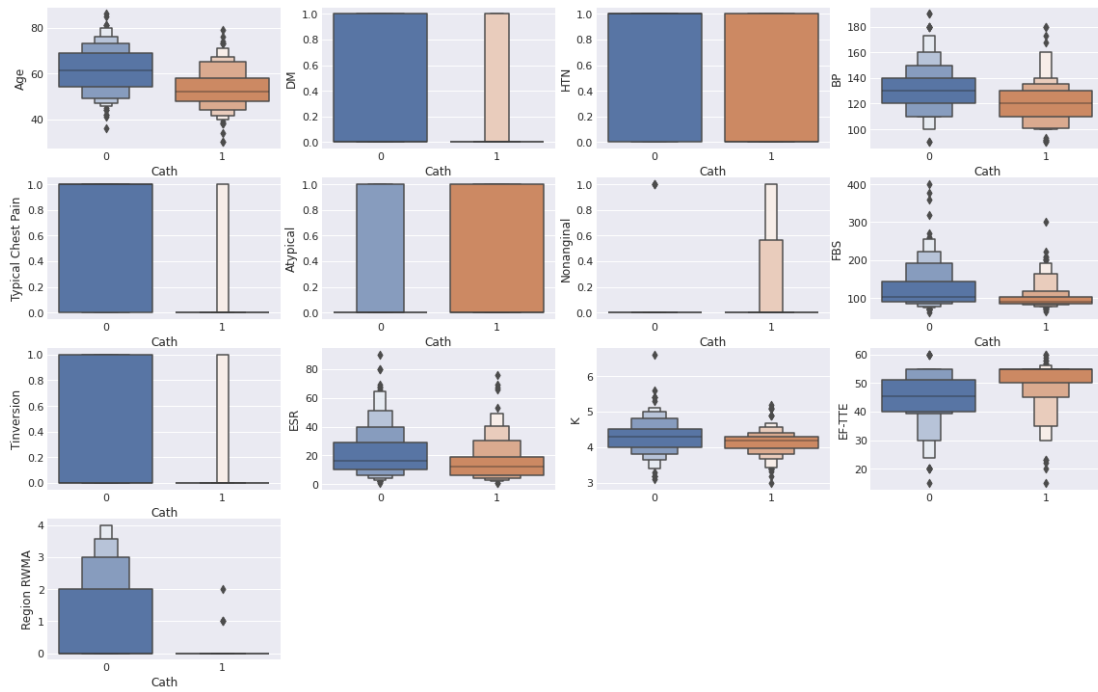


Fig. 6. The Box Plot of the Features.

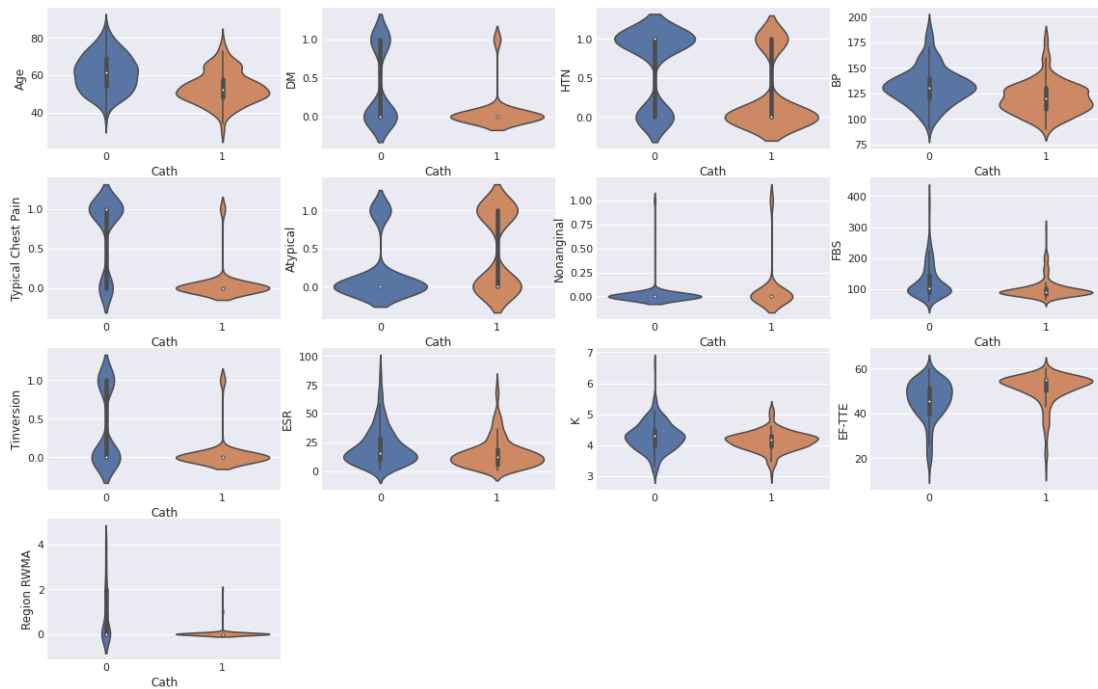


Fig. 7. The Violin Plot of the Features.

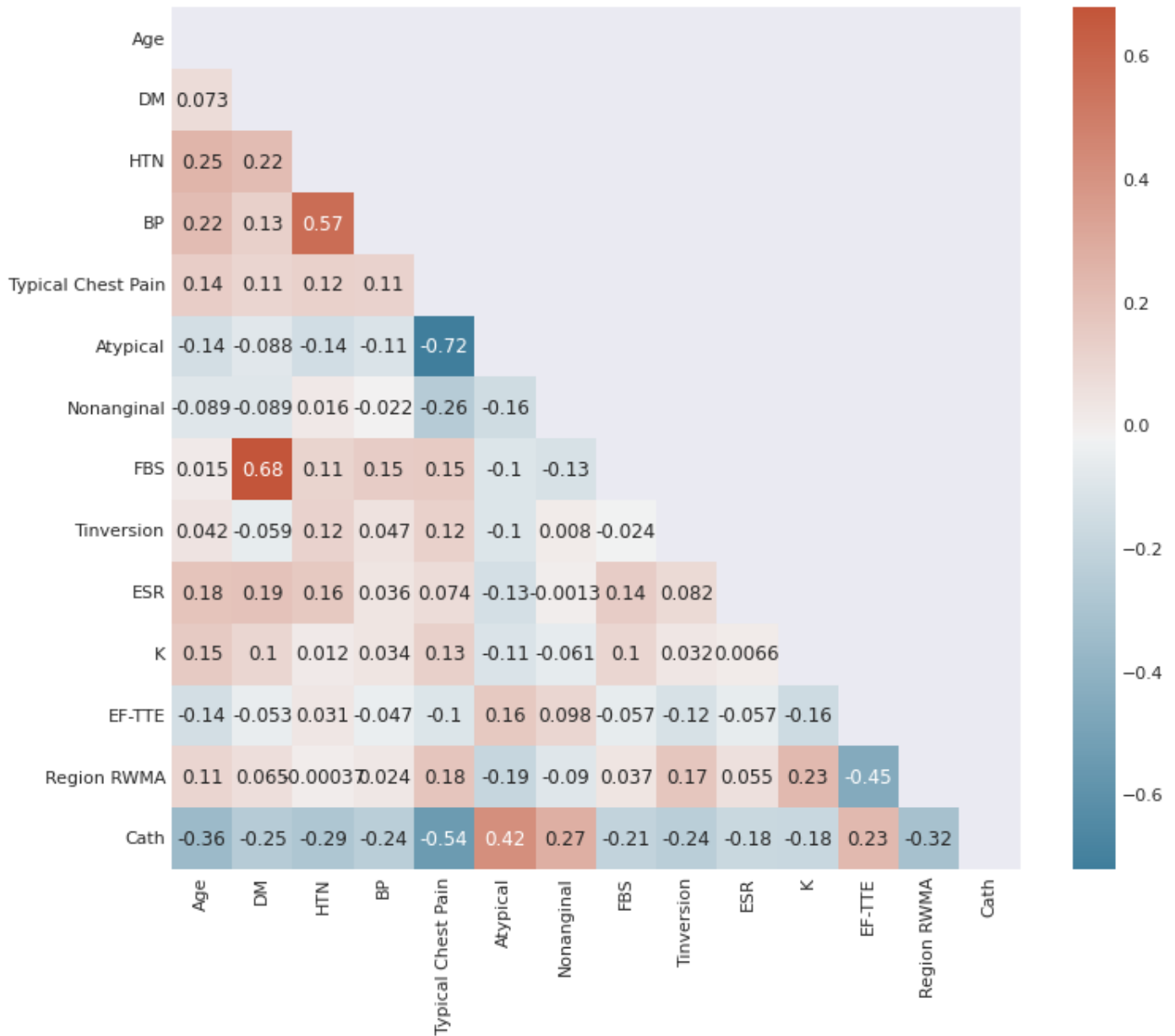


Fig. 8. Correlation Heatmap.

$Sa_1, Sa_2, \dots, Sa_k$  of roughly the same sizes The estimator will be trained  $k$  folds and then tested every time  $\eta 1, 2, 3, \dots, k$  it is trained on  $Sa_t$  and then tested again on  $Sa_t$  The accuracy of the cross validation technique is calculated as the number of classifications that are correct subdivided by the total number of records in the dataset, Thus formally we can state that  $Sa_i$  is the test dataset containing the instance  $m_i=(r_i, p_i)$  and therefore the accuracy of the cross validation is

$$accuracy_{cv} = \frac{1}{n} \sum_{(r_i, m_i) \in S} \sigma(I(Sa\xi_{(i)}, r_{je}), p_i) \quad (3)$$

2) **Confusion Matrix:** The Confusion Matrix typically assesses the outcome of the classification model for a given

request. This summarizes the count values of the correct and incorrect hypotheses by effective class. Table III illustrates the confusion matrix. For the purpose of this study, the negative class is the 0 class and the positive class is the 1 class. With True Positive (TP) showing the positives that are correctly classified and True Negative (TN) showing the negatives that are correctly classified, False Positive (FP) shows misclassified instances that are positive, and False Negative (FN) represents misclassified negative instances, respectively.

3) **Accuracy:** The accuracy of the model is the proportion of correctly classified prediction points divided by the number of total predictions evaluated, as follows:

$$Accuracy = \frac{(TP + TN)}{(TP + FN + FP + TN)} \quad (4)$$

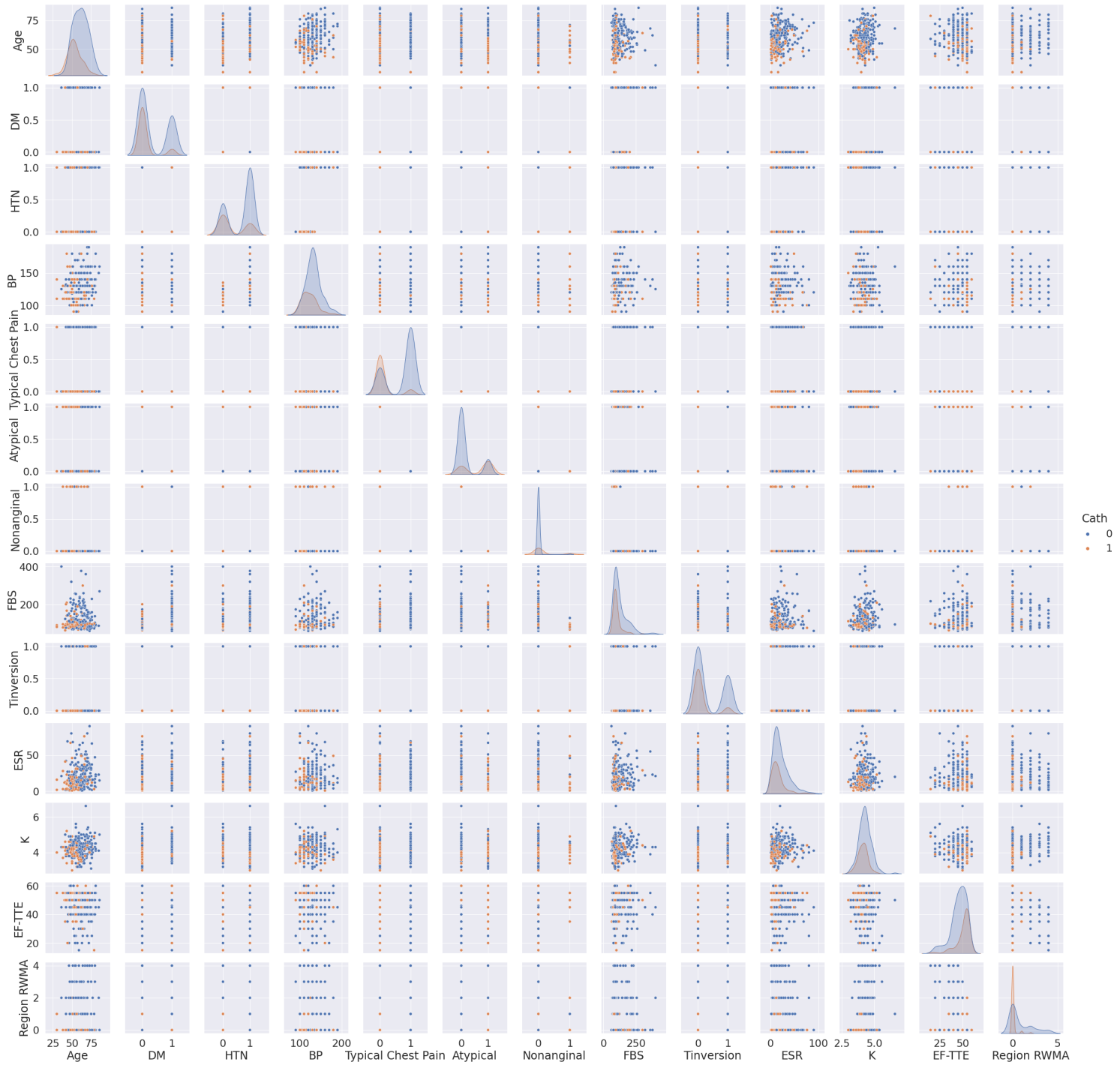


Fig. 9. The ScatterPlot Matrix of the Features.



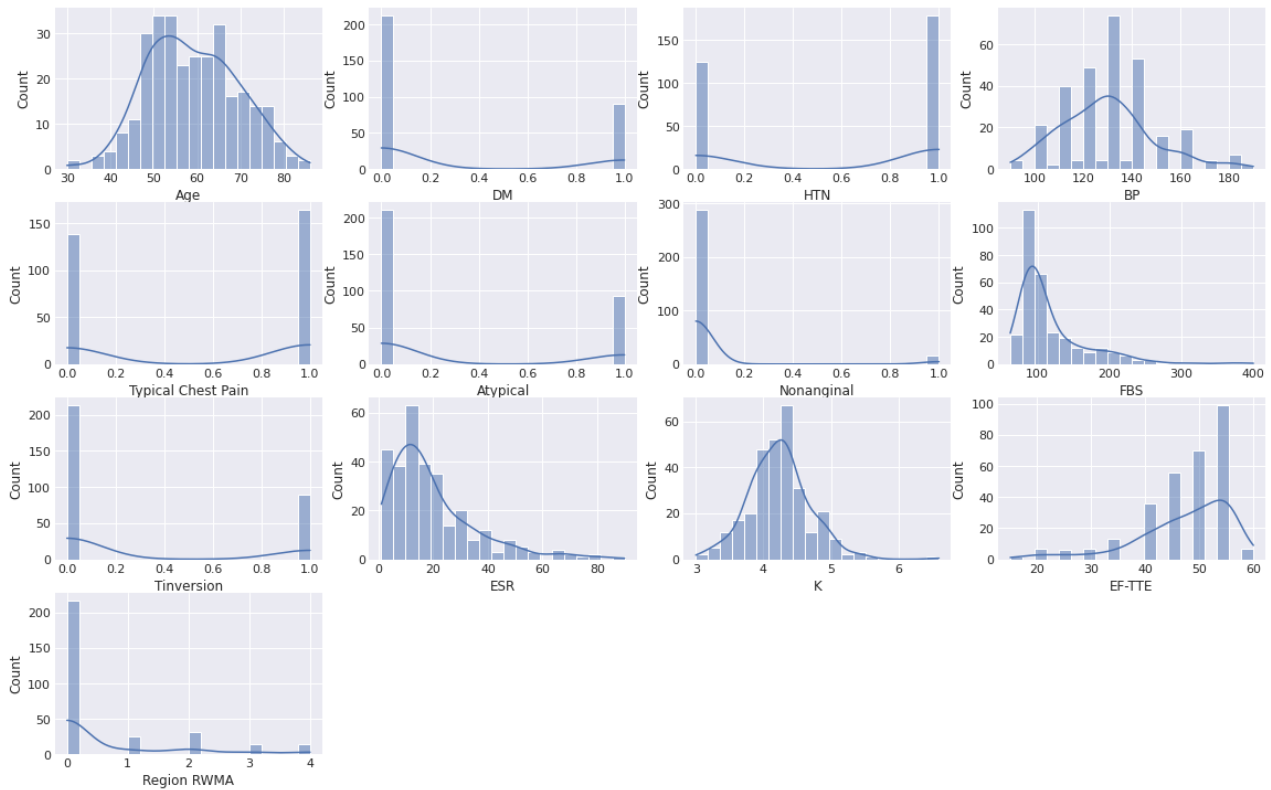


Fig. 10. Histogram of each Feature.

TABLE I. RANGE VALUE OF SELECTED FEATURES FROM THE Z-ALIZADEH SANI DATASET

Feature Type	Attribute	Values
Demographic	Age	30–86
	Diabetes Milletus(DM)	Y,N
	Hypertension(HTN)	Y,N
Symptom and examination	Blood Pressure(BP)	90–190
	Typical Chest Pain	Y,N
	Atypical	Y,N
	Nonanginal	Y,N
ECG	T inversion	0,1
Laboratory tests	Fasting Blood Sugar(FBS)	62–100 mg/dl
	Erythrocyte Sed rate(ESR)	1–90 mm/h
	Potassium(K)	3.0–6.6 mEq/lit
	Ejection Fraction(EF-TTE)	15–60%
	Regional Abnormality(Region RWMA)	0,1,2,3,4

4) *Sensitivity*: This is calculated by dividing the ratio of the number of coronary patients diagnosed as true positives by the total number of patients with coronary artery disease. It, or the true positive rate, is also called recall. It is assessed as following:

$$Recall = \frac{(TP)}{(TP + FN)} \quad (5)$$

5) *Specificity*: The specificity, or “True Negative” TN rate, is the percentage of reported diseases that are correctly

diagnosed. It is assessed as follows:

$$Specificity = \frac{(TN)}{(TN + FP)} \quad (6)$$

#### B. Results of the Machine Learning Algorithms

We implemented a variety of models and used the cross-validation technique with 10 folds, in order to select the best performing models, the more accurate models are employed in the voting ensemble, and the resulting accuracies of the models are shown in the Table IV and as it is represented

TABLE II. STATISTICAL SUMMARY OF SELECTED FEATURES FROM Z-ALIZADEH SANI DATASET

Feature	Count	Mean	Std	Min	25%	50%	75%	Max
Age	303.0	58.897690	10.392278	30.0	51.0	58.0	66.0	86.0
DM	303.0	0.297030	0.457706	0.0	0.0	0.0	1.0	1.0
HTN	303.0	0.590759	0.492507	0.0	0.0	1.0	1.0	1.0
BP	303.0	129.554455	18.938105	90.0	120.0	130.0	140.0	190.0
Typical Chest Pain	303.0	0.541254	0.499120	0.0	0.0	1.0	1.0	1.0
Atypical	303.0	0.306931	0.461983	0.0	0.0	0.0	1.0	1.0
Nonanginal	303.0	0.052805	0.224015	0.0	0.0	0.0	0.0	1.0
FBS	303.0	119.184818	52.079653	62.0	88.5	98.0	130.0	400.0
Tinversion	303.0	0.297030	0.457706	0.0	0.0	0.0	1.0	1.0
ESR	303.0	19.462046	15.936475	1.0	9.0	15.0	26.0	90.0
K	303.0	4.230693	0.458202	3.0	3.9	4.2	4.5	6.6
EF-TTE	303.0	47.231023	8.927194	15.0	45.0	50.0	55.0	60.0
Region RWMA	303.0	0.620462	1.132531	0.0	0.0	0.0	1.0	4.0
Cath	303.0	0.287129	0.453171	0.0	0.0	0.0	1.0	1.0

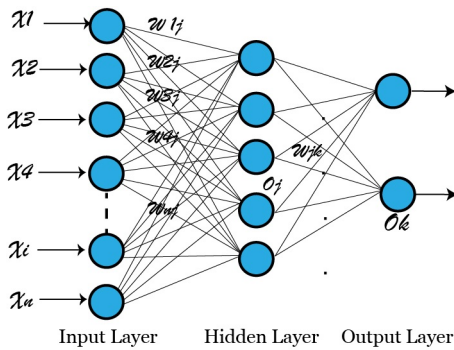


Fig. 11. The Multilayer Feed Forward Neural Network.

TABLE III. CONFUSION MATRIX

	CAD	NORMAL
Actual CAD	TP	FP
Actual Normal	FN	TN

graphically in Fig. 13, the best-performing machine learning classifiers are the RandomForest, Multilayer Perceptron, Stacking, Bagging and Adaboost. In addition, different ensembles were constructed and tested by combining these classifiers, as shown in Fig. 13 and detailed in Table IV, We evaluated the classifiers in terms of accuracy, sensitivity, specificity, F-measure, and Matthew’s correlation coefficient (MCC) to measure performance. As shown in Table IV, the ensemble voting classifier has the greatest classification accuracy of 88.12% compared to the other classifiers. Taking into account the other factors, the voting classifier has the greatest F-measure and MCC with values of 88.12 and 73.4, respectively, as illustrated in Fig. 14. The ensemble voting classifier has the best precision of 89.4% and the best recall of 88.1%, while the Multilayer Perceptron has the second-best precision of 87.79%. Once again, the voting classifiers have the best ROC and the precision values of 93.21% and 89.41% respectively, as shown in Fig. 14. The diagnostic ability of the classifier is shown in Fig. 12 by the calculated and presented ROC curves. The better the diagnostic ability of the model, the closer the ROC curve area value is to one.

### V. LIMITATIONS AND FUTURE WORK

The Z-Alizadeh Sani dataset contains the records of 303 patients from a nearby population of the Department of Cardiovascular Imaging, Rajaei Cardiovascular Medical Research Center, University of Iran, Tehran, Iran. Some limitations of the Z-Alizadeh Sani dataset are that patients under 30 years of age are not presented, as well as people from developing or low-income countries who are at high risk of developing CAD. This is to allow generalization of the proposed approach to a larger population with Coronary Artery Disease, To overcome this limitation, we suggest extending this research beyond the Z-Alizadeh Sani dataset to other CAD datasets and then investigating its generalizability to state-of-the-art machine learning models, The aim will be to design a one-time diagnostic system for Coronary Artery Disease, regardless of age or origin.

### VI. CONCLUSION

The aim of this paper is to design a more accurate classification model that predicts coronary artery disease by taking advantage of clinical and non-clinical features such as symptoms, examination, ECG, and laboratory tests. This will support remote monitoring and diagnosis of patients using vital signs and gathering features. In order to enhance the classification results with respect to accuracy, sensitivity, specificity, and Matthews correlation coefficient, however, the accuracy is improved by incorporating the gain ratio feature selection method. In addition, the benchmark dataset is experimented with to check whether there is a meaningful enhancement in prediction using the feature selection methods among the twelve classifier models. The proposed ensemble voting classifier outperforms the State-of-the-Art Machine Learning Models in terms of precision, accuracy, recall, and F-measure. The results of the proposed ensemble voting classifier are even more encouraging, as it achieved a prediction accuracy of 88.12% compared to the other classifiers. Therefore, an e-diagnosis tool based on an Ensemble Voting classifier (RF + Adaboost + MLP) would be beneficial to remote patients through cost-effective diagnosis and monitoring. Furthermore, the research can be extended by using other datasets to predict other diseases.

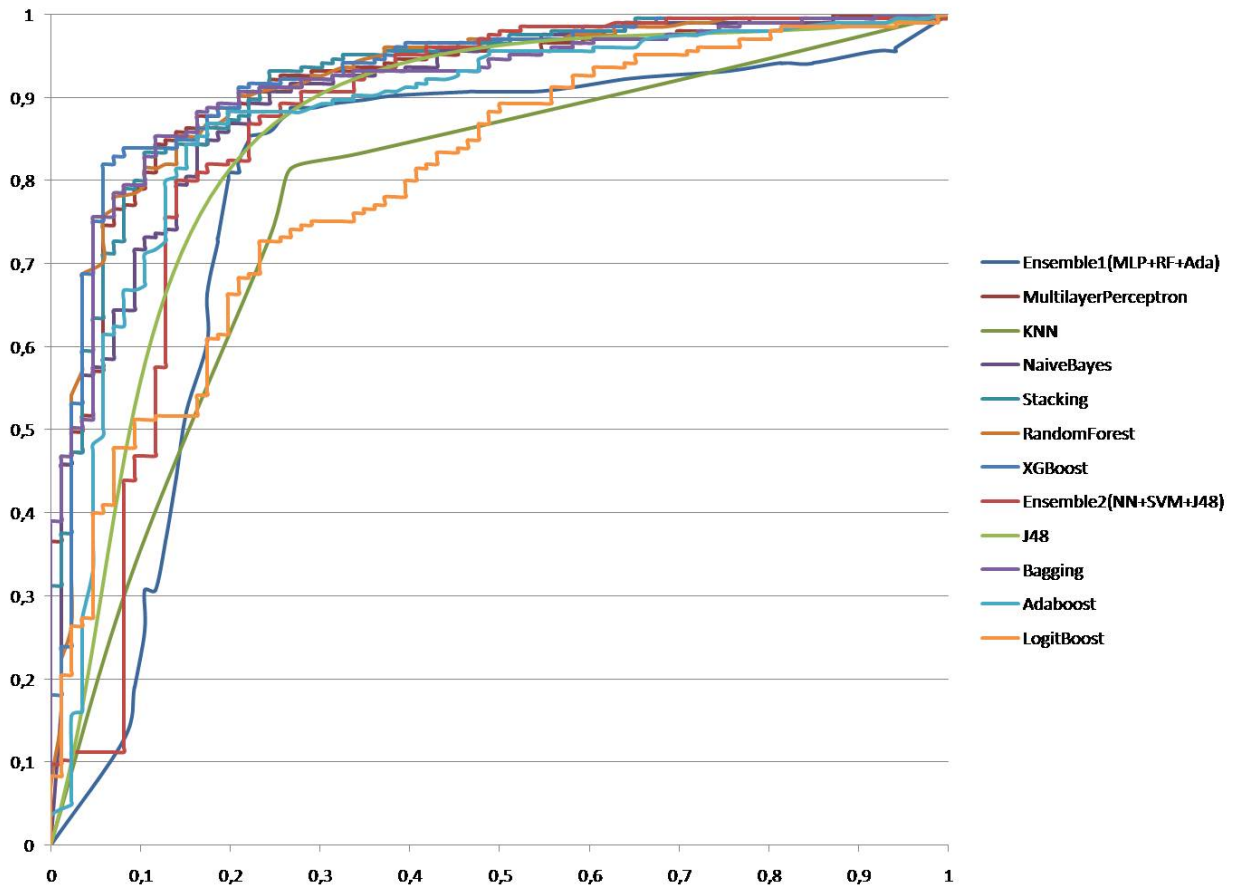


Fig. 12. Comparison of the ROC Curves of the Proposed Ensemble Voting Model with State-of-the-Art Machine Learning Models

TABLE IV. COMPARISON OF THE PROPOSED MODEL WITH STATE-OF-THE-ART MACHINE LEARNING MODELS

Model	Accuracy	Precision	Recall	F-Measure	MCC	ROC Area	Kappa	RMSE
Ensemble 1(MLP +RF+Adaboost)	88,12%	89,40%	88,10%	88,40%	0,734	0,932	0,7267	0,3137
MultilayerPerceptron	87,79%	88,00%	87,80%	87,90%	0,707	0,927	0,7067	0,3033
Stacking	87,13%	86,90%	87,10%	87,00%	0,679	0,927	0,6778	0,3101
Bagging	87,13%	88,00%	87,10%	87,40%	0,703	0,932	0,699	0,315
RandomForest	86,47%	86,40%	86,50%	86,40%	0,668	0,918	0,6683	0,3133
Ensemble2(SVM+KNN+J48)	86,47%	87,00%	86,50%	86,70%	0,681	0,85	0,6794	0,3678
AdaBoost	85,15%	85,90%	85,10%	85,40%	0,653	0,917	0,6504	0,3504
J48	84,82%	85,10%	84,80%	84,90%	0,634	0,848	0,6342	0,3634
XGboost	84,82%	85,10%	84,80%	84,90%	0,634	0,871	0,6342	0,4668
NaiveBayes	82,84%	83,60%	82,80%	83,10%	0,597	0,899	0,5947	0,3731
LogitBoost	81,19%	82,40%	81,20%	81,60%	0,567	0,827	0,563	0,4143
KNN	78,88%	79,90%	78,90%	79,30%	0,507	0,781	0,5044	0,4583

REFERENCES

[1] M. Bouma, M. Bouma, G. de Grooth *et al.*, “Standaard stabiele angina pectoris versie 4.0,” *M43: Nederlands huisartsen genootschap*, 2020.

[2] B. B. Hoorweg, R. T. Willemsen, L. E. Cleef, T. Boogaerts, F. Buntinx, J. F. Glatz, and G. J. Dinant, “Frequency of chest pain in primary care, diagnostic tests performed and final diagnoses,” *Heart*, vol. 103, no. 21, pp. 1727–1732, 2017.

[3] M. Bouma, F. Rutten, A. Bohnen, and T. Wiersma, “Samenvatting van de nhg-standaard stabiele angina pectoris (tweede herziening),” *Nederlands tijdschrift voor geneeskunde*, vol. 148, no. 45, pp. 2221–2225, 2004.

[4] C. Rudin, “Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead,” *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206–215, 2019.

[5] R. Alizadehsani, J. Habibi, M. J. Hosseini, H. Mashayekhi, R. Boghrati, A. Ghandeharioun, B. Bahadorian, and Z. A. Sani, “A data mining approach for diagnosis of coronary artery disease,” *Computer methods and programs in biomedicine*, vol. 111, no. 1, pp. 52–61, 2013.

[6] M. M. Dhanvijay and S. C. Patil, “Internet of things: A survey of enabling technologies in healthcare and its applications,” *Computer Networks*, vol. 153, pp. 113–131, 2019.

[7] F. R. Yazdi, M. Hosseinzadeh, and S. Jabbehdari, “A review of state-of-the-art on wireless body area networks,” *International Journal of Advanced Computer Science and Applications*, vol. 8, no. 11, 2017.

[8] J. Y. Khan, M. R. Yuce, G. Bulger, and B. Harding, “Wireless body area network (wban) design techniques and performance evaluation,” *Journal of medical systems*, vol. 36, no. 3, pp. 1441–1457, 2012.

- [9] D. C. Yadav and S. Pal, "Analysis of heart disease using parallel and sequential ensemble methods with feature selection techniques: heart disease prediction," *International Journal of Big Data and Analytics in Healthcare (IJBDAH)*, vol. 6, no. 1, pp. 40–56, 2021.
- [10] J. P. Li, A. U. Haq, S. U. Din, J. Khan, A. Khan, and A. Saboor, "Heart disease identification method using machine learning classification in e-healthcare," *IEEE Access*, vol. 8, pp. 107 562–107 582, 2020.
- [11] A. Javeed, S. Zhou, L. Yongjian, I. Qasim, A. Noor, and R. Nour, "An intelligent learning system based on random search algorithm and optimized random forest model for improved heart disease detection," *IEEE Access*, vol. 7, pp. 180 235–180 243, 2019.
- [12] S. Saxena, V. K. Gupta, and P. Hrisheeksha, "Coronary heart disease detection using nonlinear features and online sequential extreme learning machine," *Biomedical Engineering: Applications, Basis and Communications*, vol. 31, no. 06, p. 1950046, 2019.
- [13] A. K. Dwivedi, "Performance evaluation of different machine learning techniques for prediction of heart disease," *Neural Computing and Applications*, vol. 29, no. 10, pp. 685–693, 2018.
- [14] N. Gupta, N. Ahuja, S. Malhotra, A. Bala, and G. Kaur, "Intelligent heart disease prediction in cloud environment through ensembling," *Expert Systems*, vol. 34, no. 3, p. e12207, 2017.
- [15] L. Verma, S. Srivastava, and P. Negi, "A hybrid data mining model to predict coronary artery disease cases using non-invasive clinical data," *Journal of medical systems*, vol. 40, no. 7, pp. 1–7, 2016.
- [16] K. H. Miao, J. H. Miao, and G. J. Miao, "Diagnosing coronary heart disease using ensemble machine learning," *International Journal of Advanced Computer Science and Applications*, vol. 7, no. 10, 2016.
- [17] N. C. Long, P. Meesad, and H. Unger, "A highly accurate firefly based algorithm for heart disease prediction," *Expert Systems with Applications*, vol. 42, no. 21, pp. 8221–8231, 2015.
- [18] M. Nilashi, H. Ahmadi, A. A. Manaf, T. A. Rashid, S. Samad, L. Shahmoradi, N. Aljojo, and E. Akbari, "Coronary heart disease diagnosis through self-organizing map and fuzzy support vector machine with incremental updates," *International Journal of Fuzzy Systems*, vol. 22, no. 4, pp. 1376–1388, 2020.
- [19] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.
- [20] P. Radivojac, N. V. Chawla, A. K. Dunker, and Z. Obradovic, "Classification and knowledge discovery in protein databases," *Journal of Biomedical Informatics*, vol. 37, no. 4, pp. 224–239, 2004.
- [21] E. H. Houssein, M. Kilany, and A. E. Hassanien, "Ecg signals classification: A review," *Int.J.Intell.Eng.Inform.*, vol. 5, no. 4, pp. 376–396, 2017.
- [22] I. H. Witten, E. Frank, M. A. Hall, C. J. Pal, and M. DATA, "Practical machine learning tools and techniques," in *Data Mining*, vol. 2, no. 4, 2005.
- [23] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996.
- [24] —, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
- [25] A. Liaw and M. Wiener, "Classification and regression by randomforest," *R News*, vol. 2, no. 3, pp. 18–22, 2002.
- [26] A. Verikas, A. Gelzinis, and M. Bacauskiene, "Mining data with random forests: A survey and results of new tests," *Pattern Recognition*, vol. 44, no. 2, pp. 330–349, 2011.
- [27] R. Arora and S. Suman, "Comparative analysis of classification algorithms on different datasets using weka," *International Journal of Computer Applications*, vol. 54, no. 13, pp. 21–25, 2012.
- [28] J. Han and M. Kamber, "Data mining: Concepts and techniques," *Data Mining: Concepts and Techniques*, 2001.
- [29] M. Minsky and S. Papert, "An introduction to computational geometry," *Cambridge tiass., HIT*, vol. 479, p. 480, 1969.
- [30] N. Baba, "A new approach for finding the global minimum of error function of neural networks," *Neural Networks*, vol. 2, no. 5, pp. 367–373, 1989.
- [31] M. Stone, "Cross-validatory choice and assessment of statistical predictions," *Journal of the royal statistical society: Series B (Methodological)*, vol. 36, no. 2, pp. 111–133, 1974.

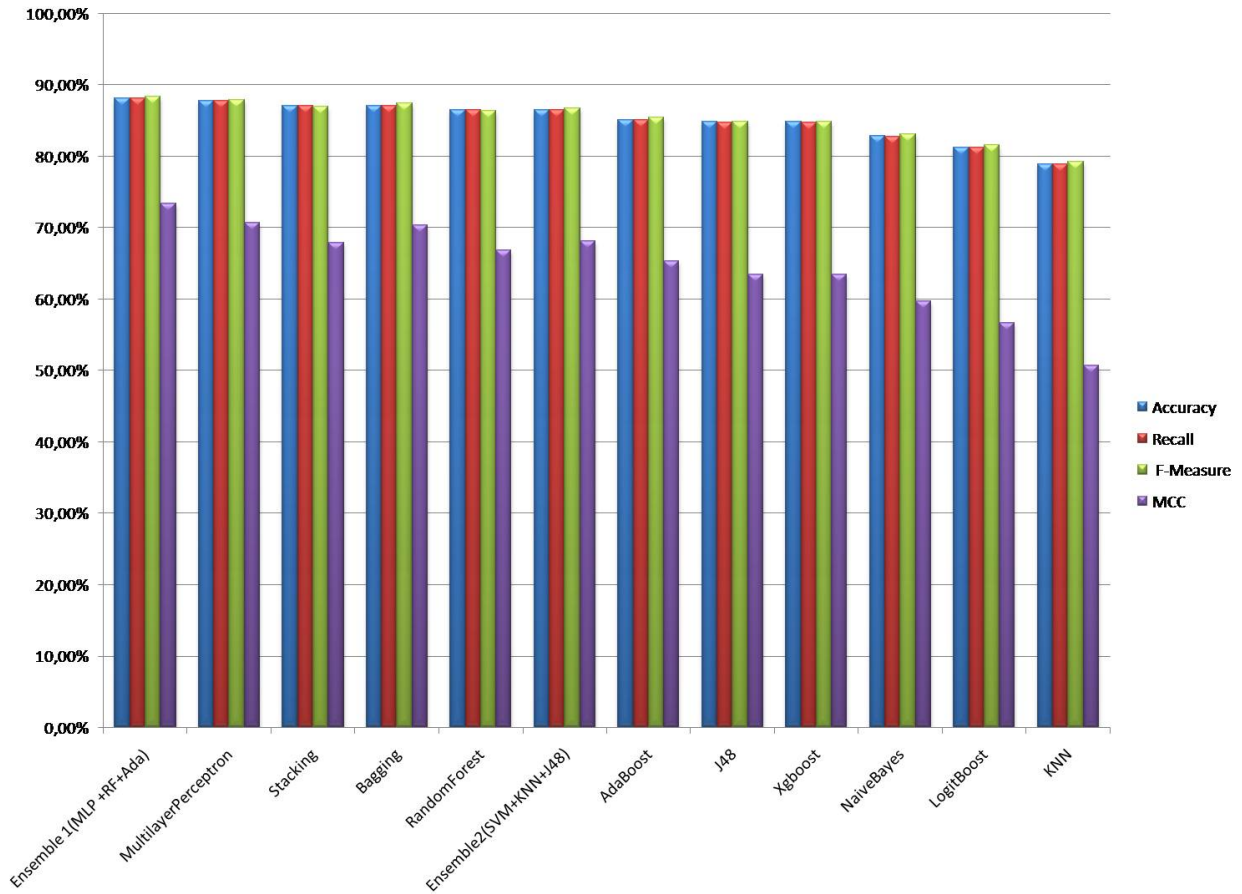


Fig. 13. Comparison of the Accuracy, Recall, F-Measure, and MCC of the Proposed Ensemble Voting Model with State-of-the-Art Machine Learning Models.

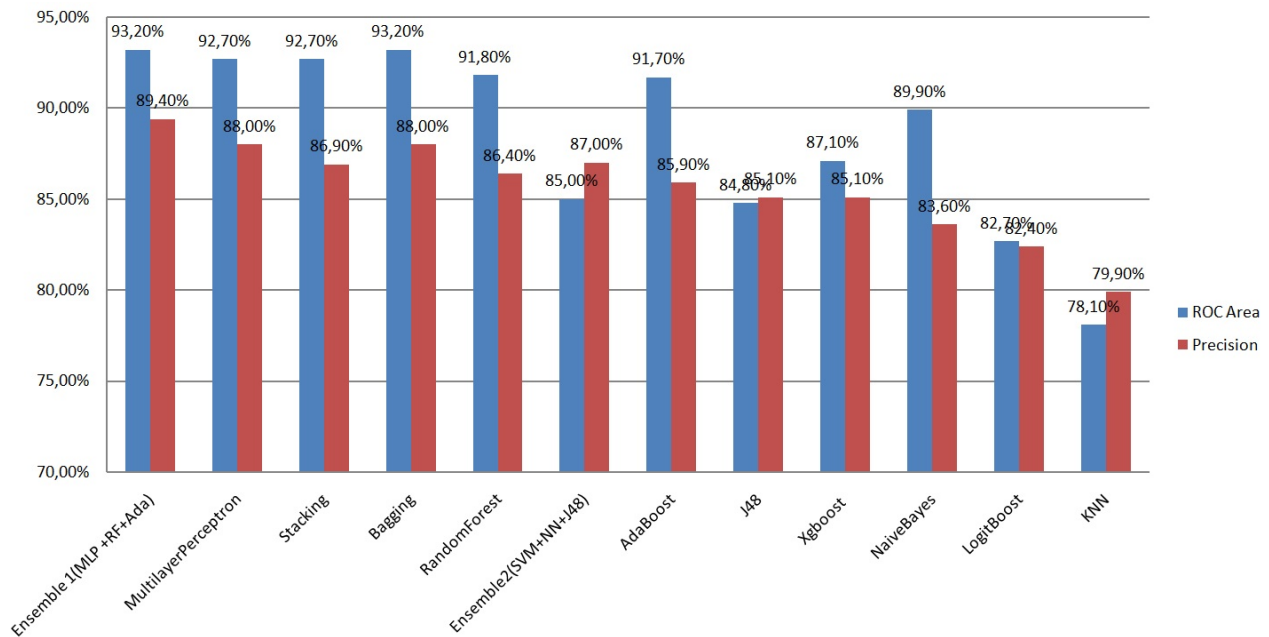


Fig. 14. Comparison of the ROC and the Precision of the Proposed Ensemble Voting Model with State-of-the-Art Machine Learning Models.

# Estimation of Varying Reaction Times with RNN and Application to Human-like Autonomous Car-following Modeling

Lijing Ma, Shiru Qu, Junxi Zhang, Xiangzhou Zhang  
School of Automation, Northwestern Polytechnical University, Xi'an, China

**Abstract**—The interaction between human-driven vehicles and autonomous vehicles has become a vital issue in micro-transportation science. Compared to autonomous vehicles, human-driven vehicles have varying reaction times that could compromise traffic efficiency and stability. But human drivers can anticipate future traffic conditions subconsciously, which guarantees qualified performance. This paper proposes an estimation method of varying reaction times and a human-like autonomous car-following model. The varying reaction times are estimated based on recurrent neural networks (RNNs) after the cross-correlation analysis of human-driven vehicles' trajectory profiles. A human-like autonomous car-following model is established based on Intelligent Driver Model (IDM), considering both varying reaction times and temporal anticipation, and the short form is IDM\_RTТА. The analytical string stability of IDM\_RTТА is deduced and illustrated. The trajectory simulation result shows that increasing accuracy of trajectory prediction is obtained with the proposed model, which will benefit the interaction between human-driven vehicles and autonomous vehicles.

**Keywords**—Car-following model; intelligent driver model; human-driven vehicle; autonomous vehicle; varying reaction time; string stability

## I. INTRODUCTION

Autonomous driving technology has developed rapidly, and with the gradually pervading use of autonomous vehicles, road traffic will experience the coexistence of human-driven and autonomous vehicles. In this situation, the interaction between human-driven vehicles and autonomous vehicles has become a vital issue in micro-transportation science. Due to the nature of human characteristics, the car-following behavior of human-driven vehicles differs from autonomous driving implemented in most microscopic models, which is a controversial topic in traffic flow analysis and simulation [1].

Reaction time is a widely recognized human driver factor, which has been incorporated into car-following modeling. Although these contributions of human-driven vehicles' reaction time have great achievement in the investigation of human characteristics and autonomous driving, the reaction time considered in car-following modeling is usually assumed as one or several constants without elaborately estimated. Furthermore, to the best of the authors' knowledge, few studies consider the inter-driver heterogeneity as well as the intra-driver heterogeneity (i.e., one driver's reaction times change from event to event [2]) of reaction times, in this case, the varying feature of reaction times is not deeply investigated and incorporated into car-following models.

To bridge these gaps, in this paper, we suggest a data-driven method to estimate the varying reaction time based on real human driving data. The estimation method includes first analyzing human-driven vehicles' trajectory profiles with cross-correlation and then learning and predicting reaction times with recurrent neural networks (RNN), which can reflect the inter-driver and intra-driver heterogeneities of reaction time. Moreover, the estimated varying reaction times are applied to car-following modeling by extending the intelligent driver model (IDM). The proposed model shows qualified analytical string stability and simulation accuracy.

The rest of this paper is organized as follows. Section II deploys the literature review of reaction time estimation and the corresponding car-following models. The estimation method for varying reaction times and the modified IDM is proposed in Section III. The estimation results are applied to the proposed car-following model, and the stability analysis and trajectory simulation are conducted, which is presented in Section IV. Section V gives the conclusion.

## II. LITERATURE REVIEW

As car-following behavior is essential to microscopic traffic research, recent studies attempt to incorporate human characteristics into car-following models for the in-depth investigation of human-driven vehicles. Compared to autonomous vehicles, human-driven vehicles, on the one hand, have reaction times that could compromise traffic efficiency and stability. On the other hand, unlike machines, human drivers can anticipate future traffic conditions subconsciously, which guarantees qualified performance. Reaction time is composed of mental processing time, body movement time, and vehicle response time [3]. Multiple studies used real trajectory data for analysis to estimate the reaction time, and there was agreement that reaction time can be estimated by the gap between stimulus and response [4], [5], [6], [7], [8]. A hybrid model is proposed by [9], which estimates the desirable acceleration of the driver by car-following model then estimates the reaction time by an artificial neural network. Reference [10] calibrated the reaction time as a constant using the field data of the intersections.

As a widely recognized human driver factor, reaction time has been incorporated into car-following modeling. The popularly used safety distance car-following model, Gipps model, involved a constant reaction time [11]. Gazis-Herman-Rothery (GHR) model [12] incorporated reaction time and was extended by [13] considering the inter-driver heterogeneity of reaction time. Optimal velocity (OV) model that takes reaction



time into account was presented in [14], and was modified by [15] with small reaction times and long reaction times considered separately for realistic performance. Reference [16] incorporated reaction time into car-following model and analyzed the impact of reaction time on traffic flow linear stability. Meanwhile, as the importance of human drivers' anticipation demonstrated by [17], the temporal anticipation was used as compensation to balance the negative effect of reaction time for stability [17], [18], [19], [20], [21].

### III. METHODOLOGY

#### A. Human-Driven Vehicles' Reaction Time

Reaction time refers to the fact that humans have an inevitable time delay in decision-making and actions, such as driving a car. Most studies presented there is a time lag from vehicle speed profile to acceleration profile and inclined to use this time lag to define the reaction time. However, some have recognized that other information in the trajectory profile is also crucial to reaction time estimation, such as gap distance [6].

Time headway is an important indicator for evaluating driving safety, closely related to traffic flow composition and driving behavior. Time headway represents the time difference between two vehicles passing through the same place. It can be calculated by dividing the headway of two vehicles (i.e., from the leading vehicle's front bumper to the subject vehicle's front bumper) by the speed of the subject vehicle, presented in (1). As time headway includes the information of gap distance as well as velocity, it can be regarded as the stimulus and acceleration as the response. Therefore, we define reaction time ( $T$ ) as the time lag between time headway and acceleration, as shown in Fig. 1.

$$thw_t = \frac{h_t}{v_t} \quad (1)$$

where  $thw_t$  and  $h_t$  are the time headway and headway between the subject vehicle and the leading vehicle at time  $t$ .  $v_t$  is the speed of the subject vehicle at time  $t$ .

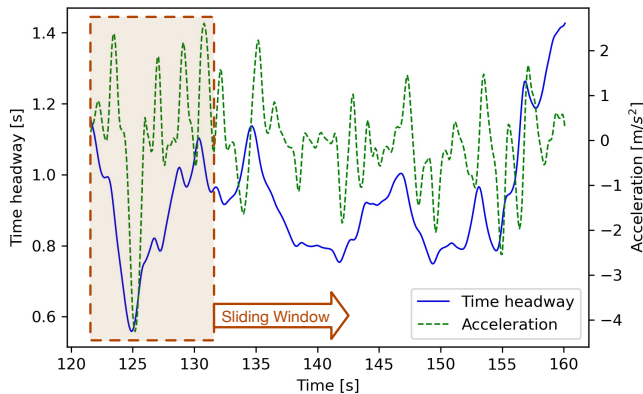


Fig. 1. The Profile of Time Headway and Acceleration.

#### B. Estimation of Varying Reaction Times based on RNN

Since it is difficult to obtain the time lags by measuring them manually, we apply the cross-correlation analysis as presented in (2)-(4) to estimate them [12]. As capturing the time-varying feature of reaction times, a sliding window is set to follow the time lags over time. To reasonably set the minima, maxima, and the length of the sliding window for estimation, we investigated the recent studies and summarized them in [22]. We assume the reaction time is distributed from 0.4 s to 3.0 s, i.e.  $\alpha = 4$  and  $\beta = 30$ , and set the length of the sliding window as 10 s, which could capture the time-varying feature of reaction time, as illustrated in Fig. 1.

$$R_\tau = E[thw_t a_{t+\tau}] \quad (2)$$

$$\rho_\tau = \frac{R_\tau - \mu_{thw}\mu_a}{\gamma_{\Delta v}\gamma_a} \quad (3)$$

$$\tau^* = \{\tau | \max(\rho_\tau), \alpha \leq \tau \leq \beta\} \quad (4)$$

where  $R_\tau$  represents the cross-correlation function.  $E[\cdot]$  denotes the expectation function.  $thw_t$  and  $a_{t+\tau}$  are the time headway and acceleration sequences.  $\mu_{thw}$  and  $\mu_a$  denote the mean values of the time headway and acceleration sequences respectively.  $\gamma_{thw}$  and  $\gamma_a$  are the standard deviations of the sequences.  $\alpha$  and  $\beta$  are the minima and maxima of the reaction time, as analyzed and assumed above.

With the cross-correlation method, we are able to collect a dataset that contains the trajectory and the corresponding reaction time. We further apply a recurrent neural network (RNN), which is good at processing sequence learning problems, to learn the varying reaction times from the dataset. Fig. 2 shows the architecture of a typical RNN that takes on an input sequence to generate an output, and the inputs are in order. The RNN learns the hidden sequence order and the corresponding output value, as presented in (5)-(6). The input  $X_t$  is the trajectory data, including gap distance ( $\Delta x_t$ ), relative speed ( $\Delta v_t$ ), speed ( $v_t$ ), acceleration ( $a_t$ ), and time headway ( $thw_t$ ). The output  $O_t$  is the reaction time ( $T$ ). Thus, the varying reaction times can be estimated from a given trajectory.

$$S_t = \tanh(U \cdot X_t + W \cdot S_{t-1}) \quad (5)$$

$$O_t = f(V \cdot S_t) \quad (6)$$

where  $U$  is the weight matrix from the input layer to the hidden layer.  $W$  is the weight matrix considering historical input data.  $V$  is the weight matrix from the hidden layer to the output layer.

#### C. Car-Following Model with Varying Reaction Times

The Intelligent Driver Model (IDM) was first proposed by [23] to simulate bottleneck congestion, as formulated in Eqs.(7)-(8). It is a distinguished mathematical car-following

model that provides collision-free behavior as well as self-organizing properties, which can be used in adaptive cruise control (ACC) [24].

$$\dot{v}_t = \tilde{a} \left[ 1 - \left( \frac{v_t}{\tilde{v}} \right)^4 - \left( \frac{S(v_t, \Delta v_t)}{\Delta x_t} \right)^2 \right] \quad (7)$$

$$S(v_t, \Delta v_t) = s_0 + t_0 v_t - \frac{v_t \Delta v_t}{2\sqrt{\tilde{a}\tilde{b}}} \quad (8)$$

where  $S(v_t, \Delta v_t)$  denotes the desired space headway function that is obtained from the vehicle's speed ( $v_t$ ) and relative speed ( $\Delta v_t$ ).  $\Delta x_t$  is the gap distance between the subject vehicle and the leading vehicle at time  $t$ , which differs from  $h_t$  and can be calculated by  $h_t - l$ , in which  $l$  is the length of the leading vehicle. The desired speed ( $\tilde{v}$ ), the maximum acceleration ( $\tilde{a}$ ), the maximum deceleration ( $\tilde{b}$ ), the desired time headway ( $t_0$ ), and the minimum space headway ( $s_0$ ) are the model parameters need to be calibrated.

We extend IDM based on varying reaction times and temporal anticipation to establish a human-like autonomous

$$\dot{v}_t = \tilde{a} \left[ 1 - \left( \frac{v_{t-T} + a_{t-T}T}{\tilde{v}} \right)^4 - \left( \frac{s_0 + t_0(v_{t-T} + a_{t-T}T) - \frac{(v_{t-T} + a_{t-T}T)(\Delta v_{t-T} - a_{t-T}T)}{2\sqrt{\tilde{a}\tilde{b}}}}{\Delta x_{t-T} + \Delta v_{t-T}T} \right)^2 \right] \quad (13)$$

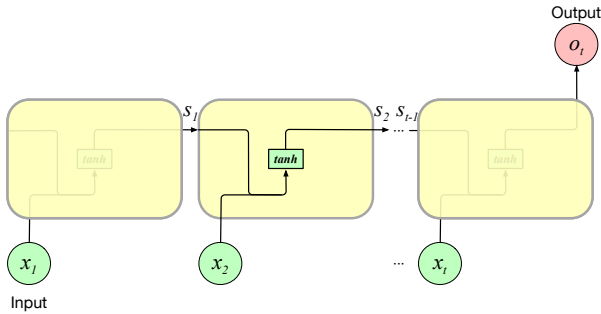


Fig. 2. RNN Architecture.

#### IV. EXPERIMENTS

##### A. Data Preparation

Analysis and simulation for human-driven vehicle interaction with autonomous vehicles require in-depth insights into human-driven vehicle behavior based on real-world human-driven vehicle trajectory data. The NGSIM dataset [25] is human-driven vehicle trajectory data collected from the real world, extracted from high-definition video by the Next Generation Simulation (NGSIM) computer program. The Interstate 80 (I-80) freeway dataset was collected on April 13, 2005, on eastbound I-80 in the San Francisco Bay Area in Emeryville, California, which was conducted under the NGSIM program. This study area has one high-occupancy lane (lane 1) and five regular lanes (from lane 2 to lane 6), including an on-ramp, and the length of this area is 503 m, as shown in Fig. 3. The author in [26] reconstructed this dataset to meet the consistency of vehicle kinematics and the reasonability of

car-following model, designated as IDM\_RT TA. Reference [17] suggested applying the constant-acceleration heuristics method to calculate future velocity. According to this concept, we formulated temporal anticipation in (9)-(11), which predict the future  $t$  timestep based on the trajectory at  $t - T$  timestep. Therefore, for the estimation result of varying reaction time ( $T$ ), the extended IDM function is established in (12), which can be further expanded as in (13).

$$\Delta x'_t = \Delta x_{t-T} + \Delta v_{t-T}T \quad (9)$$

$$\Delta v'_t = \Delta v_{t-T} - \dot{v}_{t-T}T \quad (10)$$

$$v'_t = v_{t-T} + \dot{v}_{t-T}T \quad (11)$$

$$\dot{v}_t = f(\Delta x', \Delta v', v')_t = f(\Delta x, \Delta v, v, \dot{v})_{t-T} \quad (12)$$

microscopic traffic dynamics. We extract car-following events from this reconstructed dataset to analyze the behavior of the human-driven vehicle and establish human-like autonomous driving. We obtain 1338 car-following events involving 636842 trajectory data points, and the trajectory resolution is 10Hz, i.e., the time interval between two adjacent time steps is 0.1 s. To ensure the testing process is independent, we select the car-following events in Lane 2 involving 330 vehicle pairs as testing data. Thus, 1008 vehicle pairs' car-following events in the rest of the lanes (from Lane 3 to Lane 6) are used as training data.

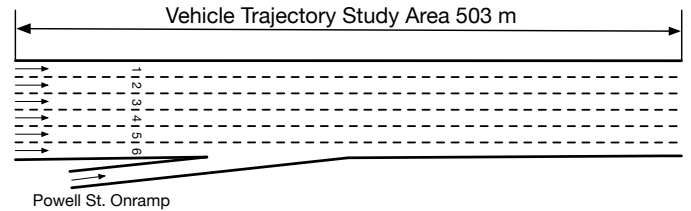


Fig. 3. Schematic of Eastbound/Northbound I-80.

##### B. Varying Reaction Times and Model Calibration

The time lags are obtained from the training dataset with the cross-correlation method and used as output data in the RNN training process. In this process, the RNN model can learn to output the reaction times from trajectory profiles, and its internal parameters update with the training samples for better output accuracy. The configuration of RNN is validated and summarized as follows:

TABLE I. CALIBRATED PARAMETERS OF IDM AND IDM\_RTТА

$T$	$\tilde{a}$	$\tilde{b}$	$\tilde{v}$	$t_0$	$s_0$
/	1.05	1.76	27.14	1.25	2.09
0.4	1.89	2.60	25.29	2.75	4.05
0.5	2.58	1.72	29.14	2.53	5.66
0.6	2.36	2.71	29.53	2.19	2.90
0.7	2.28	1.80	26.77	2.28	2.29
0.8	2.34	2.63	29.93	2.68	4.76
0.9	2.07	1.29	28.50	2.56	3.23
1.0	2.09	2.40	27.12	2.98	2.33
1.1	1.51	1.29	27.85	2.17	2.08

- Number of hidden layers: 1
- Number of neurons: 10
- The cost function: mean squared error
- The optimization algorithm: Adam [27]
- Batch size: 32
- Epochs: 20

The well-trained RNN model is applied to testing data. The estimated reaction times are obtained, and their percentages are illustrated in Fig. 4. The result indicates that this method could estimate the varying reaction times from trajectory profiles. The great majority of reaction times distribute from 0.4 s to 1.1 s. Then, this method will be integrated into the extended IDM model (IDM\_RTТА) for online estimation to imitate human-driven vehicle behavior as human-like as possible.

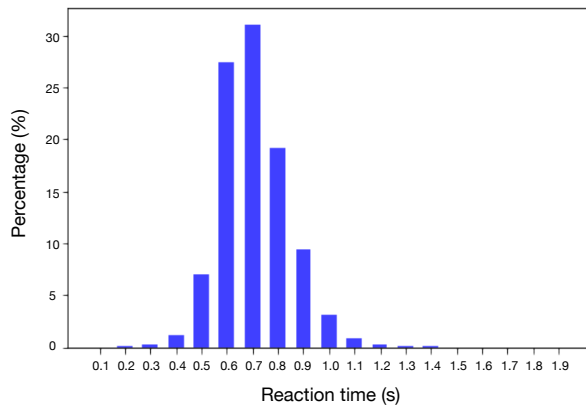


Fig. 4. Distribution of Dynamic Reaction Times.

We calibrate the proposed IDM\_RTТА for different reaction times ( $T = 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0, 1.1$ ) with the Genetic Algorithm (GA) [28], which is a stochastic global search optimization algorithm. The calibrated results are presented in Table I. The parameters of IDM are also calibrated, as shown in the first line of this table, for further model comparison.

### C. Stability Analysis

The stability of car-following models is a hot topic in the studies of traffic flow theory [29]. A car-following model can be simply represented by a control equation as in (12), and its equilibrium situation is presented in (14).

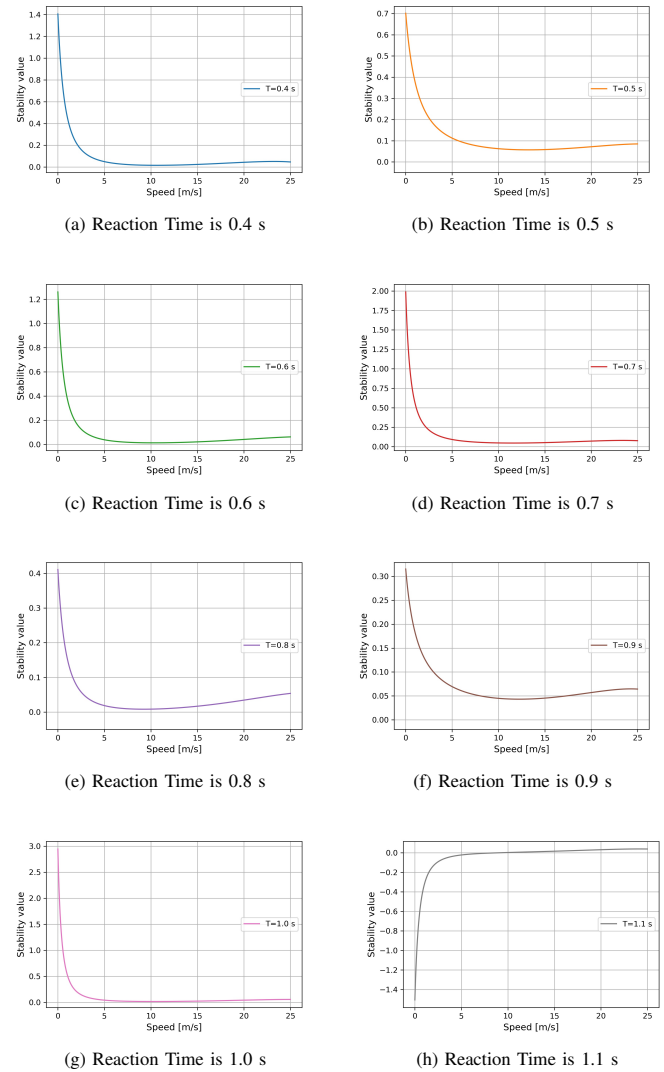


Fig. 5. String Stability for Varying Reaction Times.

$$f(\Delta x_{t-T}^*, 0, v_{t-T}^*, 0) = 0 \quad (14)$$

where acceleration and relative speed are zero in equilibrium situation, i.e.,  $\dot{v}_t = a_{t-T} = 0$  and  $\Delta v_{t-T} = 0$ , and there is an equilibrium solution for gap distance and speed, i.e.,  $\Delta x_{t-T} = \Delta x_{t-T}^*$  and  $v_{t-T} = v_{t-T}^*$ .

Empirical observations suggest that the unstable speed-spacing relationship can manifest in traffic flow as only one vehicle deviates from equilibrium (e.g., a slight change in speed) because the perturbations propagate to upstream traffic [23]. If a car-following model has string stability, for an infinite platoon of vehicles in equilibrium, the disturbance will decay as it propagates upstream [23], [30]. Although there are plenty of studies on the stability of the car-following models [1], as the IDM\_RTТА model is built with varying reaction times, the impact of this new variable on the string stability of traffic flow needs to be explored.

The research in [30] supposed a small deviation from the equilibrium state and deduced the control equations. The unstable condition are given in (15):

$$\frac{1}{2}f_v^2 - f_{\Delta v}f_v - f_{\Delta x} < 0 \quad (15)$$

where  $f_{\Delta x}, f_{\Delta v}, f_v$  are the partial differential of the coupled differential equation for gap distance, relative speed, and speed, which can be calculated by (16).

$$\begin{cases} f_{\Delta x} = \frac{\partial f(\Delta x, \Delta v, v, a)}{\partial \Delta x} \Big|_{(\Delta x_{t-T}^*, 0, v_{t-T}^*, 0)} \\ f_{\Delta v} = \frac{\partial f(\Delta x, \Delta v, v, a)}{\partial \Delta v} \Big|_{(\Delta x_{t-T}^*, 0, v_{t-T}^*, 0)} \\ f_v = \frac{\partial f(\Delta x, \Delta v, v, a)}{\partial v} \Big|_{(\Delta x_{t-T}^*, 0, v_{t-T}^*, 0)} \end{cases} \quad (16)$$

This string stability analysis method is applied to IDM\_RTTA, and the following partial differential results are derived in (17). Combining (15) and (17), the string stability of IDM\_RTTA can be evaluated, and for the varying reaction times and corresponding parameters presented in Table I, the stability value against equilibrium speed is plotted in Fig. 5. This clearly visualizes that the calibrated IDM\_RTTA with varying reaction times stays stable at most circumstances. The model is unstable for equilibrium speeds below 10 m/s when reaction time is 1.1 s, as shown in Fig. 5h.

$$\begin{aligned} f_{\Delta x} &= 2\tilde{a} \frac{(s_0 + t_0 v_{t-T}^*)^2}{\Delta x_{t-T}^{*3}} \\ f_{\Delta v} &= \sqrt{\frac{\tilde{a}}{\tilde{b}}} \frac{v_{t-T}^*(s_0 + t_0 v_{t-T}^*)}{\Delta x_{t-T}^{*2}} - 2\tilde{a} \frac{T(s_0 + t_0 v_{t-T}^*)^2}{\Delta x_{t-T}^{*3}} \\ f_v &= -2\tilde{a} \left[ \frac{2}{\tilde{v}} \left( \frac{v_{t-T}^*}{\tilde{v}} \right)^3 + \frac{t_0(s_0 + t_0 v_{t-T}^*)}{\Delta x_{t-T}^{*2}} \right] \end{aligned} \quad (17)$$

#### D. Trajectory Simulation

Trajectory simulation enables measuring the driving models' operation accuracy by comparing them to human-driven trajectories. The comparison is usually evaluated by a measure of performance (MoP), as shown in (18). The simulated trajectory (speed-time and space-time profiles) can be calculated by a discrete-time car-following process based on the acceleration obtained from IDM\_RTTA, as formulated in (19). Moreover, the mean square error (MSE) of the simulated and observed trajectories is calculated.

$$\text{MSE} = \frac{1}{M} \sum_{t=1}^M [x_t - \hat{x}_t]^2 \quad (18)$$

$$\begin{cases} \hat{v}_t = \hat{v}_{t-1} + \hat{a}_t \Delta t \\ \hat{x}_t = \hat{x}_{t-1} + \hat{v}_{t-1} \Delta t + \frac{1}{2} \hat{a}_t \Delta t^2 \end{cases} \quad (19)$$

Where  $x_t$  denotes the observed location.  $\hat{a}$  denotes the predicted acceleration.  $\hat{v}$  and  $\hat{x}$  denote the estimated speed and location.

TABLE II. MEASURE OF PERFORMANCE (MSE)

Model	Mean	SD
IDM	34.37	75.66
IDM_RTTA	29.75	57.95

Trajectories are simulated on the testing data (330 car-following events) with IDM and IDM\_RTTA. The statistical metrics of MSE are compared (see Table II), and the trajectory profiles of one randomly selected car-following event are shown in Fig. 6. IDM is a mathematical car-following model with excellent performance, and simulations show that it can provide reliable results for trajectory reproduction. Their MSE values distribute with different mean values (34.37 for IDM and 29.75 for IDM\_RTTA). Moreover, the standard deviation (SD) suggests that the IDM\_RTTA model has a smaller range of MSE values and a denser distribution pattern. These results indicate that the extended human-like car-following model not only produces higher accuracy in reproducing trajectories but also shows a more stable driving quality.

#### V. CONCLUSION

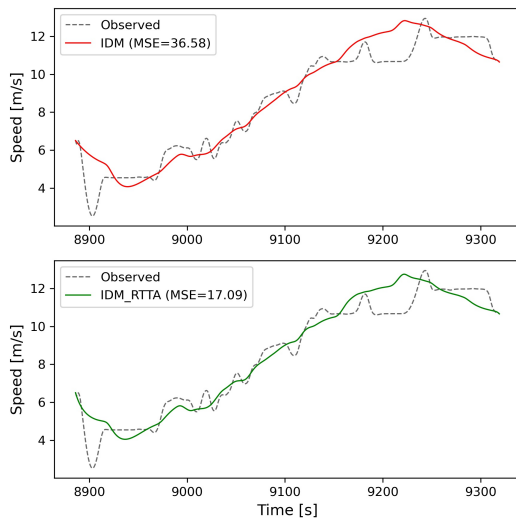
In conclusion, we focus on estimating varying reaction times and the human-like autonomous car-following modeling. We investigate the existing reaction time estimation methods and identify the importance and possibility of incorporating human-driven vehicles' time-varying reaction times and temporal anticipation properties into autonomous car-following modeling. The extended IDM (IDM\_RTTA) shows qualified string stability and demonstrates higher simulation accuracy for longitudinal control. Although the high-fidelity NGSIM data used in this paper is suitable for discovering human driver behavior, the proposed estimation method for varying reaction times still needs to be validated and tested on more real-world datasets, which is the direction of future work.

#### ACKNOWLEDGMENT

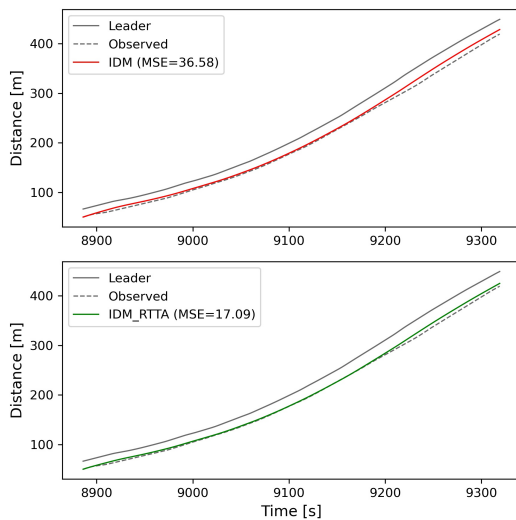
The authors would like to thank the editors and reviewers for their time and valuable insights.

#### REFERENCES

- [1] M. Saifuzzaman and Z. Zheng, "Incorporating human-factors in car-following models: a review of recent developments and research needs," *Transportation research part C: emerging technologies*, vol. 48, pp. 379–403, 2014.
- [2] T. Kim and H. Zhang, "A stochastic wave propagation model," *Transportation Research Part B: Methodological*, vol. 42, no. 7-8, pp. 619–634, 2008.
- [3] M. Green, "'how long does it take to stop?' methodological analysis of driver perception-brake times," *Transportation human factors*, vol. 2, no. 3, pp. 195–216, 2000.
- [4] R. E. Chandler, R. Herman, and E. W. Montroll, "Traffic dynamics: studies in car following," *Operations research*, vol. 6, no. 2, pp. 165–184, 1958.
- [5] X. Ma and I. Andréasson, "Driver reaction delay estimation from real data and its application in gm-type model evaluation," *Transportation Research Record*, no. 1965, pp. 130–141, 2006.
- [6] J. Zheng, K. Suzuki, and M. Fujita, "Car-following behavior with instantaneous driver-vehicle reaction delay: A neural-network-based methodology," *Transportation research part C: emerging technologies*, vol. 36, pp. 339–351, 2013.



(a) Speed-Time Profiles



(b) Space-Time Profiles

Fig. 6. Comparison of Simulated Trajectories of One Randomly Selected Car-Following Event

[7] A. Sharma, Z. Zheng, J. Kim, A. Bhaskar, and M. M. Haque, "Estimating and comparing response times in traditional and connected environments," *Transportation Research Record*, vol. 2673, no. 4, pp. 674–684, 2019.

[8] M. Zhu, X. Wang, and J. Hu, "Impact on car following behavior of a forward collision warning system with headway monitoring," *Transportation research part C: emerging technologies*, vol. 111, pp. 226–244, 2020.

[9] M. Rafati Fard, S. Rahmani, and A. Shariat Mohaymany, "Incorporating instantaneous reaction delay in car-following models: a hybrid approach," *Transportation research record*, vol. 2675, no. 10, pp. 1297–1311, 2021.

[10] A. M. Rahimi, E. Salehi, and A. Mazaheri, "Calibration of car fol-

lowing model based on two parameters related to reaction time and reaction time at stop (case study: Isfahan metropolitan)," *Journal of Transportation Research*, vol. 19, no. 3, pp. 195–210, 2022.

[11] P. G. Gipps, "A behavioural car-following model for computer simulation," *Transportation Research Part B: Methodological*, vol. 15, no. 2, pp. 105–111, 1981.

[12] D. C. Gazis, R. Herman, and R. W. Rothery, "Nonlinear follow-the-leader models of traffic flow," *Operations research*, vol. 9, no. 4, pp. 545–567, 1961.

[13] K. I. Ahmed, "Modeling drivers' acceleration and lane changing behavior," Ph.D. dissertation, Massachusetts Institute of Technology, 1999.

[14] M. Bando, K. Hasebe, K. Nakanishi, and A. Nakayama, "Analysis of optimal velocity model with explicit delay," *Physical Review E*, vol. 58, no. 5, p. 5429, 1998.

[15] L. Davis, "Modifications of the optimal velocity traffic model to include delay due to driver reaction time," *Physica A: Statistical Mechanics and its Applications*, vol. 319, pp. 557–567, 2003.

[16] Z. Yao, T. Xu, Y. Jiang, and R. Hu, "Linear stability analysis of heterogeneous traffic flow considering degradations of connected automated vehicles and reaction time," *Physica A: Statistical Mechanics and its Applications*, vol. 561, p. 125218, 2021.

[17] M. Treiber, A. Kesting, and D. Helbing, "Delays, inaccuracies and anticipation in microscopic traffic models," *Physica A: Statistical Mechanics and its Applications*, vol. 360, no. 1, pp. 71–88, 2006.

[18] A. Kesting and M. Treiber, "Calibrating car-following models by using trajectory data: Methodological study," *Transportation Research Record*, vol. 2088, no. 1, pp. 148–156, 2008.

[19] D. Sun, D. Chen, M. Zhao, W. Liu, and L. Zheng, "Linear stability and nonlinear analyses of traffic waves for the general nonlinear car-following model with multi-time delays," *Physica A: Statistical Mechanics and its Applications*, vol. 501, pp. 293–307, 2018.

[20] A. Jafaripournimchahi, W. Hu, and L. Sun, "An asymmetric-anticipation car-following model in the era of autonomous-connected and human-driving vehicles," *Journal of advanced transportation*, vol. 2020, 2020.

[21] V. Kurtc, I. Anufriev, and D. Trufanov, "Multi-anticipative car-following model with explicit reaction-time delay," *Mathematical Models and Computer Simulations*, vol. 13, no. 6, pp. 1109–1115, 2021.

[22] L. Ma and S. Qu, "A sequence to sequence learning based car-following model for multi-step predictions considering reaction delay," *Transportation research part C: emerging technologies*, vol. 120, p. 102785, 2020.

[23] M. Treiber, A. Hennecke, and D. Helbing, "Congested traffic states in empirical observations and microscopic simulations," *Physical review E*, vol. 62, no. 2, p. 1805, 2000.

[24] A. Kesting, M. Treiber, M. Schönhof, and D. Helbing, "Adaptive cruise control design for active congestion avoidance," *Transportation Research Part C: Emerging Technologies*, vol. 16, no. 6, pp. 668–683, 2008.

[25] FHWA, "The Next Generation Simulation (NGSIM) [Online]," Available: <http://www.ngsim.fhwa.dot.gov/>, 2008.

[26] M. Montanino and V. Punzo, "Trajectory data reconstruction and simulation-based validation against macroscopic traffic patterns," *Transportation Research Part B: Methodological*, vol. 80, pp. 82–106, 2015.

[27] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[28] M. J. Kochenderfer and T. A. Wheeler, *Algorithms for optimization*. Mit Press, 2019.

[29] A. Talebpour and H. S. Mahmassani, "Influence of connected and autonomous vehicles on traffic flow stability and throughput," *Transportation Research Part C: Emerging Technologies*, vol. 71, pp. 143–163, 2016.

[30] J. A. Ward, "Heterogeneity, lane-changing and instability in traffic: A mathematical approach," Ph.D. dissertation, University of Bristol Bristol, UK, 2009.

# Mobile Food Journalling Application with Convolutional Neural Network and Transfer Learning: A Case for Diabetes Management in Malaysia

Jason Thomas Chew<sup>1</sup>, Patrick Hang Hui Then<sup>4</sup>

School of Information and  
Communication Technologies  
Swinburne University of Technology  
Sarawak Campus, Malaysia<sup>1,4</sup>

Yakub Sebastian<sup>2</sup>

College of Engineering, IT & Environment  
Charles Darwin University, Australia

Valliapan Raman<sup>3</sup>

Department of AI and DS  
Coimbatore Institute of Technology  
Coimbatore, India

**Abstract**—Diabetes is an ever worsening problem in modern society, placing a heavy burden on healthcare systems. Due to the association between obesity and diabetes, food journaling mobile applications are an effective approach for managing and improving the outcome of diabetics. Due to the efficacy of nutritional tracking and management in managing diabetes, we implemented a deep learning-based Convolutional Neural Network food classification model to aid with food logging. The model is trained on a subset of the Food-101 and Malaysian Food 11 datasets, including web-scraped images, with a focus on food items found locally in Malaysia. In our experiments, we explore how fine-tuning of the image dataset improves the performance of the model.

**Keywords**—Convolutional neural network; deep learning; diabetes; food journal; mobile application; nutritional tracking; Malaysia

## I. INTRODUCTION

Diabetes is becoming an increasingly prevalent disease in modern society. It was estimated that in 2017 at least 6% of world population were affected by type 2 diabetes (T2DM) [1]. Urbanization has caused a shift in dietary patterns among the populace towards food with a higher associated risk of diabetes [2]. Diet affects an individual's diabetes outcome in various ways, both directly and indirectly. The sugar content of a meal directly influences an individual's blood glucose levels, whereas an indirect factor such as the consumption of high-fat foods causes obesity, which then contributes to the onset of diabetes [3]. It is well-established that there is a strong link between obesity and diabetes, where excess weight plays a part in up to 90% of the cases of T2DM [4]. Consequently, diet management plays a crucial role in managing the disease outcome among the diabetics and in disease prevention among the non-diabetics.

In Malaysia, the incidence rate of diabetes in adults has increased from 15.2% in 2011 [5] to 17.5% in 2015 [6]. A long-

standing health problem for the country, diabetes in Malaysia lays an increasingly heavy burden on the country's healthcare system, costing it approximately RM2.04 billion in 2011 alone [7]. Due to the association between diet and diabetes [8], food journaling applications can improve T2DM disease outcomes by helping users to plan and monitor their diet. The use of food journaling software and other related nutritional tracking applications is an emerging approach that may improve an individual TD2M management outcome [9], [10], [11]. For instance, *Diabetes Notepad* was developed to assist diabetics in managing the disease via self-care through diet monitoring, leading to an improved clinical outcome of diabetes in Korea [12]. In Malaysia, despite a consensus that having mobile food journaling app with image-snapping feature would be highly desirable to help with T2DM management, more than half of the people surveyed were unaware of such dietary logging applications [13].

In this paper, we present a new food journaling application tailored to the Malaysian demography. The challenges in developing the application are non-trivial. Because the onset T2DM is strongly linked to environmental, sociodemographic and cultural factors [14], an effective food journaling application should be contextualised to its target population and capable of recognising nutritional components of local diets (e.g. *Diabetes Notepad* is targeted at South Korean demography). Many food databases still contain inconsistent or missing nutritional content of certain food makes it difficult to estimate one's nutritional intake [9]. For instance, they may not sufficiently document non-standard food types such as restaurant food, ethnic food, food prepared by friends and party food. It is therefore important to develop a solution which focuses on a specific demography such as Malaysia.

Our contributions are two-fold. First, we propose a new mobile food journalling application that uses a convolutional neural network (CNN) architecture to enhance the classifica-



tion of food images taken with mobile phones. The amount of calorie intake from the classified food is automatically calculated for the users to help with personal diet planning and monitoring, which is crucial in managing such chronic diseases as T2DM. Second, we propose a way to enhance the performance of the classification model via transfer learning, a strategy that appears to work well on non-conventional images of local Malaysian food.

The remaining of this paper is organized as follows. Section II reviews some related work, with a specific appreciation of food journalling app development in Malaysian context. Section III describes the proposed system architecture, software components, and the method for training CNN classifier. In Section IV, we explain our experiments and report the results. Section V summarises and discusses the experimental results. The limitations of the system are described in Section VI. Section VII concludes this paper.

## II. RELATED WORK

### A. Food Recognition and Classification

Food recognition and classification algorithms form the internal machinery of food journalling apps, executed as part of the food entry workflow that reduce the burden of food entry by users. Convolutional Neural Networks (CNNs) are a popular method for food classification due to their increased efficacy over traditional machine learning-based approaches [15]. The success of CNNs can be attributed to transfer learning, where a model that is pre-trained on a large-scale image dataset such as ImageNet is retrained on a different dataset [16], [17].

There are several existing works in the literature that apply transfer learning to train CNN-based food classification models [15], [18], [19], [20], [21]. Meyers et al. [19] developed *Im2Calories*, a food diary application that uses an ensemble of CNN classifiers for food detection, classification, segmentation, and volume estimation. Their system is able to detect food from 23 different restaurants as well as perform general food detection. Pan et al. [20] proposed *DeepFood* which leverages CNN-based transfer learning algorithms for deep feature extraction. The author's goal is to classify fresh food ingredient images indexed in the Mealcome image dataset [22]. In the most recent example, Sahoo et al. [21] developed *FoodAI*, a CNN model trained to classify various dishes with a focus on food found locally in Singapore. The model is able to predict a wide variety of Singaporean and South-East Asian dishes, recognizing 756 distinct food classes.

### B. Food Journalling Apps Development in Malaysia

There is limited documentation of food journalling apps that specifically target T2DM management for Malaysian population. A recent survey of mobile diabetes management apps showed that existing apps focused only on reporting and setting reminders for meal time, without automatic food recognition ability [23]. Where automatic food image recognition is available, it is not in the context of T2DM management. For instance, [24] used artificial neural network to recognise five different types of Malaysian traditional dessert: *curry puff*, *'kuih ketayap'*, *'kuih kosui'*, *red tortoise cake*, and *'putu piring'*. Calories were subsequently computed from the detected food images but no implication of T2DM management was

documented. The algorithm reported 80% food recognition accuracy. [13] proposed a food journalling app that uses deep neural network model trained with InceptionV3 and MobileNetV2 on Food-101 dataset. Although the app was designed to assist T2DM management, it does not provide the capability to automatically compute calorie amount based on classified images.

## III. METHOD

In this section, we describe the internal workings of our proposed food journalling app. Fig. 1 overviews the essential workflow of the app. The user snaps a picture of food item, which is sent to a web server running the food image classification algorithm. The classification result is subsequently returned and displayed to the user, who then confirms the food type and provides further information about the amount of food that was consumed. The app relays these information back to the server, which computes and returns the estimated consumed calories. Fig. 2 shows the technology components used for the development of the app.

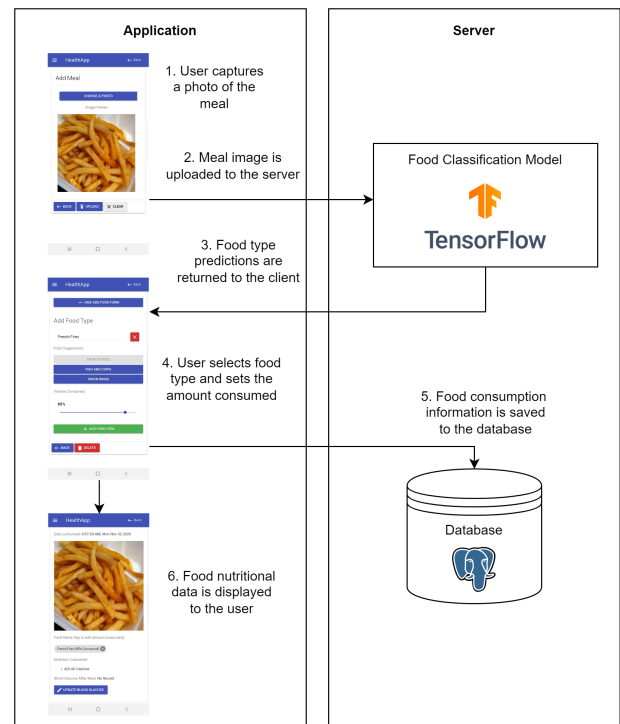


Fig. 1. Workflow of the Proposed Food Journalling App.

### A. Food Dataset

The food image classifier forms the backbone of the app. To construct an initial dataset for training the model, we combined images from the Food-101 [25] and Malaysian Food 11<sup>1</sup> datasets. Several food categories regularly consumed by Malaysians have previously been identified by [26]. These include cereal products, beverages, fruits and vegetables, confectioneries, meat products, fish products, milk products,

<sup>1</sup><https://www.kaggle.com/karkengchan/malaysia-food-11>

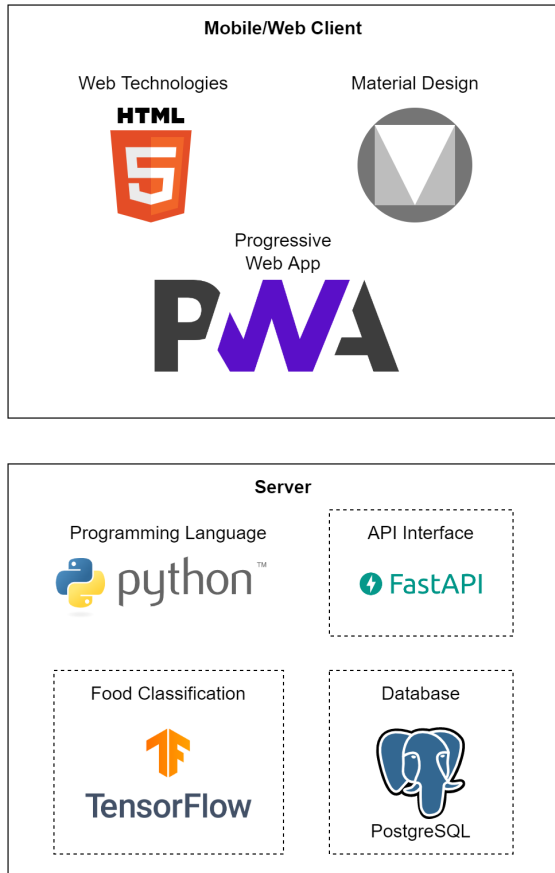


Fig. 2. Technology Components Used by the Proposed App.

condiments, eggs, legumes, and spreads. We use these categories to select food classes from the two datasets for training the model. An additional local noodle dish from Sarawak, *kolo mee*, is included in the dataset. The *kolo mee* food images were scraped from the web.

In cases where both the Food-101 and Malaysian Food 11 datasets have overlapping food classes (*fish and chips*, *fried rice* and *hamburger*), the images from the Malaysian food dataset are used for training the model and the Food-101 images are discarded. This is to reduce the class imbalance in the dataset while also ensuring that the training images are as visually similar as possible to the way the local dishes are prepared. The *apple pie* and *kolo mee* classes were manually processed for reasons explained later. One image was removed from the *Nasi Lemak* class due to corruption. The remaining food classes were used without any further preprocessing. The final dataset consists of 31335 images across 32 classes. Details of the food dataset for training the model are shown in Table I.

**B. Model Training**

We implemented the CNN using the *Keras* deep learning framework with *TensorFlow 2.2* backend and trained it on the dataset. To shorten the training duration and improve the model’s performance, transfer learning is used for training the model. Transfer learning allows for the features learnt in

TABLE I. FOOD CLASSES SELECTED FOR TRAINING THE MODEL

Food Class	Images	Food Categories
Apple Pie	883	Confectionaries
Caesar Salad	1000	Fruits & Vegetables
Chocolate Cake	1000	Confectionaries
Donuts	1000	Confectionaries
Dumplings	1000	Meat, Fruits & Vegetables
Fish and Chips	1000	Fish
French Fries	1000	Fruits & Vegetables
Fried Noodles	1000	Cereal
Fried Rice	1000	Cereal
Garlic Bread	1000	Cereal
Hamburger	1000	Meat
Hot Dog	1000	Meat
Ice Cream	1000	Confectionaries
Kaya Toast	1000	Cereal, Condiments
Kolo Mee	473	Cereal, Meat
Laksa	1000	Cereal
Lasagna	1000	Cereal
Mixed Rice	1000	Cereal, Meat, Fruits & Vegetables
Nasi Lemak	999	Cereal
Onion Rings	1000	Fruits & Vegetables
Pancakes	1000	Cereal
Peking Duck	1000	Meat
Pizza	1000	Cereal, Meat
Popiah	1000	Meat, Fruits & Vegetables
Ramen	1000	Cereal
Roti Canai	1000	Cereal
Satay	1000	Meat
Spaghetti Bolognese	1000	Cereal, Meat
Spaghetti Carbonara	1000	Cereal, Meat
Steak	1000	Meat
Tiramisu	1000	Confectionaries
Waffles	1000	Confectionaries

one domain to be transferred into another domain, such that the resulting performance on the new domain is improved, as opposed to training a new model from scratch [27], [28].

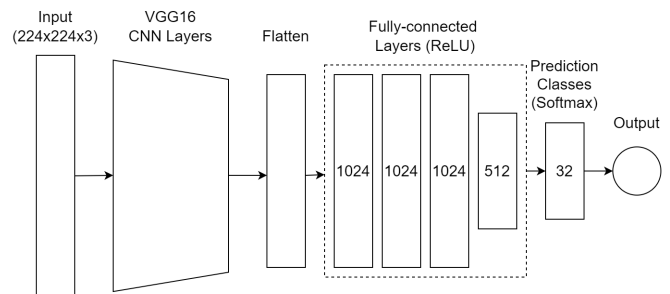


Fig. 3. CNN and Transfer Learning Architecture

Fig. 3 illustrates our CNN and transfer learning training architecture. The base model used for transfer learning is VGG16 [29], initialized with weights pre-trained on the ImageNet dataset [17]. The weights of the convolutional layers were frozen so that they do not update, while the fully connected layer of the model was replaced with three fully-connected layers with widths of 1024, 1024, 1024, and 512, respectively, and the output layer was also updated to 32 classes.

The images in the dataset are split into three: 60% for training, 20% for validation and 20% for testing. In order to prevent the model from overfitting, the images in the training set were augmented. Image augmentation allows the model to generalize better by applying a random transformation on the images according to certain parameters as they are being

fed into the model during training [30]. The parameters used for augmentation of the test images that produced the best performing model are the following: image rotation between  $-90$  and  $90$  degrees, random horizontal flipping of the images, varied brightness ranging between  $75\%$  to  $100\%$ , and image size rescaling between  $80\%$  to  $100\%$ .

The model was trained on a local machine equipped with an Nvidia GTX 1060 6GB GPU. Training was performed in three steps with different learning rates and a fixed batch size of 50. The initial learning rate was set to  $1 \times 10^{-4}$  and the model was trained until the accuracy stopped increasing. When the accuracy stopped increasing, the learning rate was updated and the training step was resumed. The learning rate for the second and third training steps were  $1 \times 10^{-5}$  and  $1 \times 10^{-6}$  respectively. A complete model takes an average of 6 hours to train. The best performing model achieved a top-1 accuracy of  $76.77\%$  and a top-5 accuracy of  $94.37\%$ .

#### IV. EXPERIMENT AND RESULTS

##### A. Initial Results

During the initial training of the model, the original apple pie images from the Food-101 dataset (1000 images) were used without any manual preprocessing. However, during testing of the model, it was found to have performed poorly in predicting apple pie images, with an accuracy of  $53.5\%$ .

Upon inspection of the apple pie images, we hypothesized that the model's below average performance was due to a portion the images in apple pie class that poorly represented the features of the food class. In order to confirm this hypothesis, we retrained the model after manually preprocessing the images to remove images that were suspected to be causing the model's poor performance. Three categories of images were identified and removed from the apple pie image class: non-food images, incorrectly labelled images, and images that poorly represent the food class, which includes images where the apple pie is mostly blocked from view or are mixed with different food classes. Fig. 4 shows a sample of the removed images. A total of 117 images were removed from the dataset and the model was retrained. The images in the other 31 classes of the dataset remain unchanged.

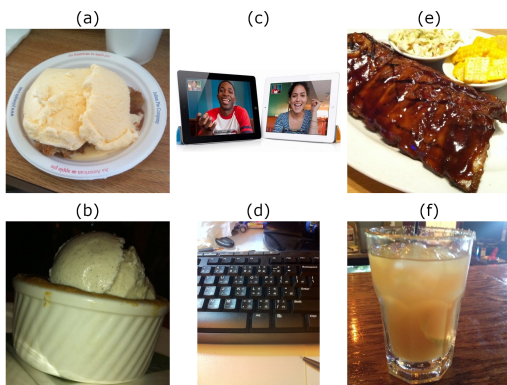


Fig. 4. The Images Removed During Preprocessing of the Apple Pie Class Includes Severely Occluded Images (a, b), Non-Food Images (c, d), and Other Food Images (e, f).

TABLE II. IMAGES MISCLASSIFIED BY THE MODEL WHEN TRAINED WITH THE UNPROCESSED AND PREPROCESSED DATASETS. THE FOOD CLASSES WITH FEWER MISCLASSIFICATIONS ARE HIGHLIGHTED.

Class	Unprocessed	Preprocessed
All Food Classes	93	72
Caesar Salad	1	1
Chocolate Cake	4	3
Donuts	5	3
Dumplings	3	2
Fish and Chips	2	2
French Fries	0	0
Fried Noodles	0	0
Fried Rice	1	3
Garlic Bread	6	3
Hamburger	1	1
Hot Dog	5	6
Ice Cream	7	4
Kaya Toast	8	4
Kolo Mee	1	0
Laksa	0	0
Lasagna	2	5
Mixed Rice	0	2
Nasi Lemak	3	0
Onion Rings	2	3
Pancakes	12	4
Peking Duck	6	3
Pizza	3	1
Popiah	0	5
Ramen	0	0
Roti Canai	3	7
Satay	2	0
Spaghetti Bolognese	0	0
Spaghetti Carbonara	1	0
Steak	0	0
Tiramisu	10	8
Waffles	5	2

After retraining using the preprocessed apple pie class, the model's performance on apple pie images noticeably improved from  $53.5\%$  to  $59.32\%$ . Table II shows the incorrect classifications of the two models, where *Unprocessed* denotes the model trained on the original 1000 *apple pie* images and *Preprocessed* denotes the model trained on the preprocessed apple pie images. There is an overall reduction in misclassifications observed in the retrained model. In the table, the model which made fewer misclassifications on a specific food class is highlighted. Several of the classes with the largest improvements observed are those which share visual similarities (*pancakes* and *kaya toast*), and *ice cream*, which is commonly served as a topping or alongside an *apple pie*.

Additionally, the retrained model was found to have misclassified fewer non-*apple pie* images as an *apple pie* despite the images in the other food classes remaining unchanged. The total number of images from the other food images which were incorrectly classified as *apple pie* was reduced from 93 to 72. These results suggest that the model's accuracy can be further improved by manually processing the images in the remaining food classes to ensure that each image in the dataset significantly represents the features of their respective classes.

##### B. Effects from Incorporating Additional Local Food Classes

Seven additional food classes were included to the dataset in order to increase the model's coverage of local food classes, namely: *Ais Kacang*, *Ayam Pansuh*, *Beef Noodle Soup*, *Boiled Eggs*, *Fried Eggs*, *Kampua Noodles*, and *Layer Cake*. The new images were obtained by scraping the web and manually processing the images to remove unrelated images. We trained two models on the new dataset, one with the same parameters

TABLE III. PREDICTION ACCURACY FOR THE MODELS TRAINED ON THE UPDATED DATASET WITH AND WITHOUT FOCAL LOSS (FL). THE COLUMNS WITH HIGHER ACCURACY ARE HIGHLIGHTED

Food Class	Images	Accuracy (No FL)	Accuracy (FL)
Ais Kacang	435	80.46%	72.41%
Ayam Pansuh	120	58.33%	66.67%
Beef Noodle Soup	520	73.08%	76.92%
Boiled Eggs	623	87.20%	85.60%
Fried Eggs	398	81.25%	77.50%
Kampua Noodles	292	63.79%	63.79%
Layer Cake	669	91.04%	86.57%
Model Accuracy	34412	75.51%	74.17%

as the original model, and another using focal loss optimizer [31]. The focal loss optimizer is introduced in an attempt to account for the high class imbalance in the new food classes. Focal loss puts greater focus on the classes that the model has difficulty classifying during the training process. The images contained in each class and the results of the two models are shown in Table III.

Upon evaluation of the two new models, the model trained using the same parameters as described in Section III-B achieved an accuracy of 75.51% whereas the model using the focal loss optimizer achieved an accuracy of 74.17%. The lower accuracy of the two new models can be attributed to the class imbalance of the newer classes due to a lack of available images on the web. The results of both models are shown in Table III. The model with focal loss is observed to perform better in classifying *Ayam Pansuh*, which has the greatest class imbalance of all the food classes. Although focal loss is observed to allow for higher performance on the highly imbalanced class, the model performs poorer in classifying the other food classes in the dataset.

### C. Sample Use Case

In this section, we describe the meal recording process on the web application. In order for users to use the application, they must first register an account and log in.

Fig. 5(a) and 5(b) shows the account creation and log in interfaces. After the user has logged in to the application, they can begin tracking their meals and view their meal history via the *Smart Diet Watcher* module. The module is accessible from the application’s main menu as shown in Fig. 5(c).

The first step in the meal logging process is uploading an image of the meal. Fig. 5(d) shows the meal creation interface for uploading a meal image. The user can choose between taking a photo of the meal or choosing an image from the device. The meal image that the user has selected is previewed before upload.

After uploading the meal image, the user can add food items to the meal. Fig. 6(a) shows the interface for adding food items to the meal. The food item suggestions are provided by the model’s predictions based on the food image. Users can add multiple food items to the meal in cases where the food image contains multiple dishes. Once food items are added to the meal, users can edit or delete them as shown in Fig. 6(b). The user can optionally record their blood glucose levels after the meal.

Once the user has completed logging their meal, they can view a summary of the recorded meal as shown in Fig. 6(c).

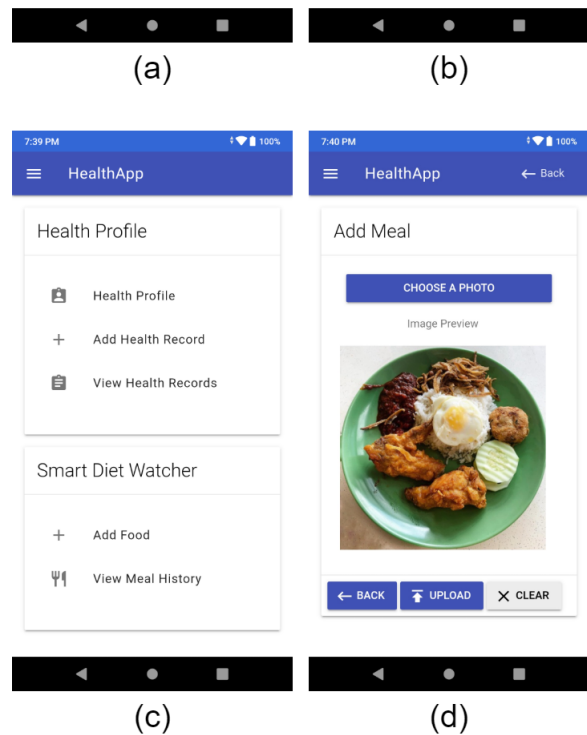
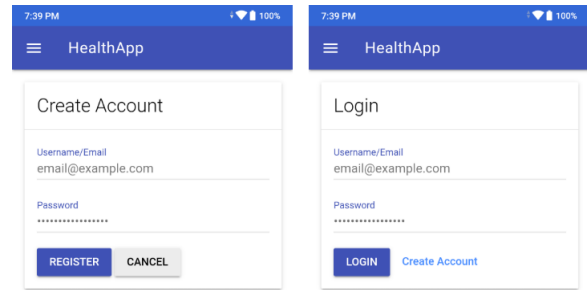


Fig. 5. (a) Shows the Account Creation, (b) Shows the Login Interface, (c) Shows the Main Menu Interface, and (d) Shows the Meal Creation Interface.

The meal summary shows the food items contained in the meal image, the meal’s nutritional value and the user’s blood glucose level after the meal. The nutritional values shown in the meal summary is calculated based on the percentage of each consumed food item that was set by the user. The baseline nutritional values that are used for the nutrition calculations are based on the amount of a single full portion serving. Lastly, the user can view their past meal records from the meal history menu as shown in Fig. 6(d).

## V. DISCUSSION

To summarise, we trained a CNN model to predict food items found locally in Sarawak, Malaysia. Two publicly avail-



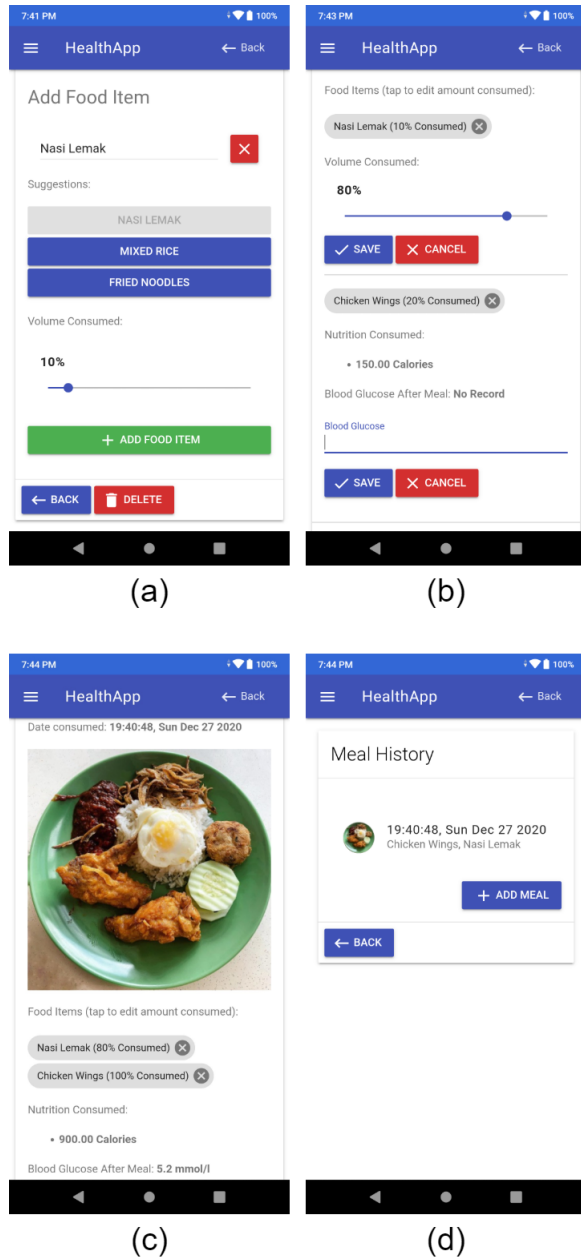


Fig. 6. (a) Shows the Interface for Adding Food Items, (b) Shows the Interface for Updating Food Items and the Post-Meal Blood Glucose Record, (c) Shows a Summary of the Meal, and (d) Shows the Meal History Menu.

able food datasets, Food-101 and Malaysian Food 11 datasets, were used for the initial model training. This was followed by incorporating local food items that were scraped from the web.

Upon training the initial model, we found that the *apple pie* food class was poorly performing. Thus, we investigated the effect of further preprocessing the images to improve the model's performance. We removed poor quality and unrelated images from the dataset, retrained the model and reported the effects of the preprocessing step in detail, shown in Table II. There was an improvement in accuracy despite the smaller number of training images, which is likely due to the images in the preprocessed dataset better representing the apple pie

class.

Next, we incorporated images of local Sarawakian food items scraped from the web into the dataset and retrained the model. Due to concerns of class imbalances for the web scraped images, focal loss was introduced. We explored the effect of focal loss on the model's performance, as shown in Table III. Interestingly, we found that the model without focal loss performed better overall, achieving a higher accuracy on most of the web scraped food classes despite the class imbalance. One notable case however is the *Ayam Pansuh* class which is severely imbalanced. The results obtained indicates that while focal loss is generally expected to improve performance when there is a class imbalance, it may be beneficial to train another model that does not implement focal loss as well in case the use of focal loss unexpectedly results in a poorer performing model.

## VI. LIMITATIONS

Currently, the number of food classes that the model can predict is limited. We plan to expand the total number of food classes that the model can predict by collecting local food images and using them to retrain the model. Our final goal is for the model to recognize most local food consumed in Malaysia.

The model's performance has been shown to improve through manual preprocessing of the apple pie class, indicating that there is a possible performance gain if all the images were manually processed. In particular, *FoodAI* [21] managed to achieve a top-1 accuracy of 83.2% where the authors put significant effort into manual inspection of the food images by domain experts while constructing the dataset. However, in our case, time and resource constraints limit the manual preprocessing of the dataset. This limitation can be addressed in future work by looking into methods of automating the dataset preprocessing via CNN-based feature extraction and clustering [32] to group visually similar images for batch labeling.

## VII. CONCLUSION

In this paper we implement a food classification model with a CNN-based architecture. Transfer learning was applied to boost the model's performance and reduce training time. We explored the effect of manually preprocessing of the dataset on the model's performance. Results indicate that the model's performance can be improved further by carefully curating the images that are used for training the model. In future work, we plan to expand the food classes that the model can predict to classify more local Malaysian dishes.

## ACKNOWLEDGMENT

The authors extend appreciation to collaborators in the Clinical Research Centre, Sarawak General Hospital for validating our works and testing our applications. The project is funded under Prototype Research Grant Scheme from the Malaysia Ministry of Higher Education, Ref: PRGS/1/2019/ICT02/SWIN/01/1.

REFERENCES

- [1] M. A. B. Khan, M. J. Hashim, J. K. King, R. D. Govender, H. Mustafa, and J. Al Kaabi, "Epidemiology of type 2 diabetes—global burden of disease and forecasted trends," *Journal of epidemiology and global health*, vol. 10, no. 1, p. 107, 2020.
- [2] V. Mohan, V. Ruchi, R. Gayathri, M. R. Bai, V. Sudha, R. M. Anjana, and R. Pradeepa, "Slowing the diabetes epidemic in the world health organization south-east asia region: the role of diet and physical activity," *WHO South-East Asia Journal of Public Health*, vol. 5, no. 1, pp. 5–16, 2016.
- [3] R. J. Johnson, L. G. Sánchez-Lozada, P. Andrews, and M. A. Lanaspá, "Perspective: a historical and scientific perspective of sugar and its relation with obesity and diabetes," *Advances in Nutrition*, vol. 8, no. 3, pp. 412–422, 2017.
- [4] S. Verma and M. E. Hussain, "Obesity and diabetes: an update," *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, vol. 11, no. 1, pp. 73–79, 2017.
- [5] Institute for Public Health Malaysia, "National health and morbidity survey 2011 (nhms 2011)," *Vol. II: Non-Communicable Diseases, Risk Factors & Other Health Problems*, 2015.
- [6] Ministry of Health Malaysia, "National health & morbidity survey 2015 non-communicable diseases, risk factors & other health problems," 2015.
- [7] F. I. Mustapha, S. Azmi, M. R. A. Manaf, Z. Hussein, J. Mahir, F. Ismail, A. N. Aizuddin, and A. Goh, "What are the direct medical costs of managing type 2 diabetes mellitus in malaysia," *Med J Malaysia*, vol. 72, no. 5, pp. 271–277, 2017.
- [8] W. Sami, T. Ansari, N. S. Butt, and M. R. Ab Hamid, "Effect of diet on type 2 diabetes mellitus: A review," *International journal of health sciences*, vol. 11, no. 2, p. 65, 2017.
- [9] F. Cordeiro, D. A. Epstein, E. Thomaz, E. Bales, A. K. Jagannathan, G. D. Abowd, and J. Fogarty, "Barriers and negative nudges: Exploring challenges in food journaling," in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 2015, pp. 1159–1162.
- [10] A. Darby, M. W. Strum, E. Holmes, and J. Gatwood, "A review of nutritional tracking mobile applications for diabetes patient use," *Diabetes technology & therapeutics*, vol. 18, no. 3, pp. 200–212, 2016.
- [11] J. Jung, L. Wellard-Cole, C. Cai, I. Koprinska, K. Yacef, M. Allman-Farinelli, and J. Kay, "Foundations for systematic evaluation and benchmarking of a mobile food logger in a large-scale nutrition study," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 4, no. 2, pp. 1–25, 2020.
- [12] Y. J. Kim, S. Y. Rhee, J. K. Byun, S. Y. Park, S. M. Hong, S. O. Chin, S. Chon, S. Oh, J.-t. Woo, S. W. Kim *et al.*, "A smartphone application significantly improved diabetes self-care activities with high user satisfaction," *Diabetes & metabolism journal*, vol. 39, no. 3, pp. 207–217, 2015.
- [13] K. K. Chan and N. Suki, "Mobile food management and dietary management for T2DM patients in malaysia," *International Journal of Scientific & Technology Research*, vol. 8, no. 11, pp. 574–578, 2019.
- [14] T. Dendup, X. Feng, S. Clingan, and T. Astell-Burt, "Environmental risk factors for developing type 2 diabetes mellitus: a systematic review," *International journal of environmental research and public health*, vol. 15, no. 1, p. 78, 2018.
- [15] V. Bruno and C. J. Silva Resende, "A survey on automated food monitoring and dietary management systems," *Journal of health & medical informatics*, vol. 8, no. 3, 2017.
- [16] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" in *Advances in neural information processing systems*, 2014, pp. 3320–3328.
- [17] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International journal of computer vision*, vol. 115, no. 3, pp. 211–252, 2015.
- [18] G. Ciocca, P. Napolitano, and R. Schettini, "Cnn-based features for retrieval and classification of food images," *Computer Vision and Image Understanding*, vol. 176, pp. 70–77, 2018.
- [19] A. Meyers, N. Johnston, V. Rathod, A. Korattikara, A. Gorban, N. Silberman, S. Guadarrama, G. Papandreou, J. Huang, and K. P. Murphy, "Im2calories: towards an automated mobile vision food diary," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1233–1241.
- [20] L. Pan, S. Pouyanfar, H. Chen, J. Qin, and S.-C. Chen, "Deepfood: Automatic multi-class classification of food ingredients using deep learning," in *2017 IEEE 3rd international conference on collaboration and internet computing (CIC)*. IEEE, 2017, pp. 181–189.
- [21] D. Sahoo, W. Hao, S. Ke, W. Xiongwei, H. Le, P. Achananuparp, E.-P. Lim, and S. C. Hoi, "Foodai: Food image recognition via deep learning for smart food logging," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019, pp. 2260–2268.
- [22] H. Chen, J. Xu, G. Xiao, Q. Wu, and S. Zhang, "Fast auto-clean cnn model for online prediction of food materials," *Journal of Parallel and Distributed Computing*, vol. 117, pp. 218–227, 2018.
- [23] S. Izahar, Q. Y. Lean, M. A. Hameed, M. K. Murugiah, R. P. Patel, Y. M. Al-Worafi, T. W. Wong, and L. C. Ming, "Content analysis of mobile health applications on diabetes mellitus," *Frontiers in Endocrinology*, vol. 8, p. 318, 2017.
- [24] N. A. A. N. Muhammad, C. P. Lee, K. M. Lim, and S. F. A. Razak, "Malaysian food recognition and calorie counter application," in *2017 IEEE 15th Student Conference on Research and Development (SCORED)*. IEEE, 2017, pp. 445–450.
- [25] L. Bossard, M. Guillaumin, and L. Van Gool, "Food-101 – mining discriminative components with random forests," in *European Conference on Computer Vision*, 2014.
- [26] I. Pondor, W. Y. Gan, and G. Appannah, "Higher dietary cost is associated with higher diet quality: a cross-sectional study among selected malaysian adults," *Nutrients*, vol. 9, no. 9, p. 1028, 2017.
- [27] L. Shao, F. Zhu, and X. Li, "Transfer learning for visual categorization: A survey," *IEEE transactions on neural networks and learning systems*, vol. 26, no. 5, pp. 1019–1034, 2014.
- [28] C. Tan, F. Sun, T. Kong, W. Zhang, C. Yang, and C. Liu, "A survey on deep transfer learning," in *International conference on artificial neural networks*. Springer, 2018, pp. 270–279.
- [29] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
- [30] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of Big Data*, vol. 6, no. 1, p. 60, 2019.
- [31] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.
- [32] J. Yang, D. Parikh, and D. Batra, "Joint unsupervised learning of deep representations and image clusters," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 5147–5156.



# Rethinking Classification of Oriented Object Detection in Aerial Images

Phuc Nguyen\*, Thang Truong\*, Nguyen D. Vo, Khang Nguyen\*\*  
University of Information Technology, Ho Chi Minh City, Vietnam  
Vietnam National University, Ho Chi Minh City, Vietnam

**Abstract**—With the help of the rapid development of technology, especially the prevalence of UAVs (unmanned aerial vehicles), object detection in aerial images gains much more attention in computer vision and deep learning. However, traditional methods use horizontal bounding boxes for object representation leading to inconsistency between objects and features. Therefore, many detectors are being built to tackle this problem, and normally they use the conventional approaches of training and testing to achieve the results. Our pipeline proposed to strengthen not only the classification but also localization via independent training processes using convex-hull transformation in data pre-processing phase. We experimented with the well-designed S2ANet, R3Det, ReDet, RoI Transformer and Oriented R-CNN on the well-established oriented object detection dataset DOTA. Then we adopt the best detectors with the well-known classification network EfficientNet to our proposed pipeline and achieve promising results on the oriented object detection DOTA dataset. Moreover, our pipeline can flexibly be adapted to various oriented object detection baselines improving the results in classification via independent extensive training cycles.

**Keywords**—Aerial images; classification; convex-hull transformation; data processing; oriented object detection

## I. INTRODUCTION

Object Detection in Aerial Images (ODAI) has always been important in our real life with tremendous real-world applications such as surveillance, disaster prediction, emergency rescue, and even urban management [1] [2] [3]. Nowadays, it is becoming more feasible thanks to the increasing growth of studies in deep learning and the fast-paced development of information and communication technology. However, objects collected from aerial images appear in a variety of representations. They are often distributed in arbitrary orientations leading to confusion for many latest deep learning models, which opens a new study aspect in computer vision. To tackle this problem, a lot of experiments conducted show that using oriented bounding box representation (OBB) instead of horizontal bounding box (HBB) representation will alleviate the mismatch features and increase object detection accuracy [4] [5] [6] [7].

ODAI is extremely crucial to this world so it requires high accuracy and fewest mismatched objects in the prediction task as possible [8]. Although detecting objects in aerial images is vital, there is a lack of data about it. Many well-designed methods follow the traditional pipeline without refining class

labels for output predictions, which could lead the model to behave biased toward less significant objects. Our proposed pipeline ensures that classes are treated equally, and models learn as much as data features from not only inside but also outside of the dataset through an independent classification training process.

In this study, we propose and provide a deep analysis of an effective training and testing pipeline to surge the performance of oriented object detectors in aerial images. Our pipeline applies the convex-hull transformation on ground-truth oriented bounding boxes to extract proper instances for the training and testing processes. Furthermore, we can use extra data for the independent training process to ensure the model classify proper label instances. We conduct extensive experiments on multiple baselines and apply the pipeline on them, yielding promising results.

We summarize our contributions in this paper as:

- Proposing a novel training and testing pipeline to improve classification performance flexibly adapt to many latest models.
- Providing a wise way to prepare data for classification training and an effective ensemble method in testing models.
- Using convex-hull transformation technique to transform oriented bounding boxes to horizontal ones for the further training process.
- Give a deep analysis of why we are choosing this pipeline and what are common problems of nowadays object detection methods in aerial images.
- Carrying out extensive experiments with the latest oriented object detection methods and providing an in-depth evaluation of the best deep learning models to strengthen our proposal.

The rest of the paper is: Section II is Related works; Section III is Methodology; Section IV is Our approach; Section V is Experiment and finally the last one, Section VI is Conclusion and Future Work.

## II. RELATED WORKS

### A. Oriented Object Detection

For the past decade, there have been various well-established object detector methods designed for tackling the horizontal object detection task, many of which have made remarkable progress such as Fast R-CNN [9], Faster R-CNN [10],

\*\*Corresponding author

\*Equal Contributor

Dynamic R-CNN [11], Deformable DETR [12], SSD [13], YOLOF [14], YOLOX [15], etc. However, these general object detection methods cannot tightly locate the object leading to the inconsistency between the classification and localization processes. Therefore, the extended branch of study using an oriented bounding box to represent the object's ground truths receives extensive attention to meet the need for applying deep learning to real-world applications (Fig. 1).

Current oriented object detection methods heavily depend on the original horizontal object detection task. They adopt the mechanism from extracting deep features and generating proposals to refining the final bounding boxes results. For example, Ding et al. introduced RoI Transformer [16] to tackle the problem of misaligned between regions of feature and objects, they applied the spatial transformation to Regions of Interest and configured the model to learn these geometric parameters using oriented bounding boxes labels. Jiang et al. introduced Rotational Region CNN [17] for detecting arbitrary-oriented texts in natural scene images. The model adopted the Faster R-CNN baseline using the region proposal network (RPN) to generate HBBs of texts and those HBBs will integrate many pooled RoI features to produce the final regressed OBBs. Han et al. introduced Single-shot Alignment Network (S2ANet) [18], addressing the issues of inconsistency between classification and localization. S2ANet consists of two main components: Feature Alignment Module (FAM) and Oriented Detection Module (ODM). The FAM specifically uses the Alignment Convolution to generate well-qualified anchors on which the active rotating filters of the ODM apply to encode the orientation information. Eventually, the network produces orientation-sensitive and orientation-invariant features to mitigate the inconsistency between classification scores and localization accuracy.

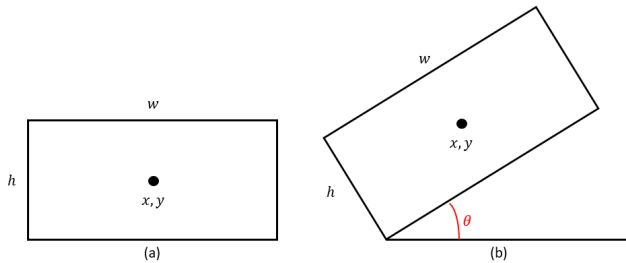


Fig. 1. Oriented Bounding Box (OBB)  $[(x, y), w, h, \theta]$ , where  $(x, y)$  is the Center and  $w, h, \theta$  are the Width, Height and Angle of an OBB.

## B. Classification

Classifying objects in aerial images using their CNN extracted features has always been a challenging problem since aerial images usually contain a tremendous number of various-shape instances. In 2020, Dosovitskiy et al. proposed a method Vision Transformer [19] following the baseline of the original architecture Transformer with the fewest modification possible. ViT splits images into many patches and connects these patches (NLP-vibes) using a sequence of linear embeddings. As the result, these patches are treated as tokens like in NLP and receive object queries for output labels. Tan and Le introduced EfficientNet [20] achieving better accuracy and efficiency than previous ConvNets by leveraging a multi-objective neural

architecture search that optimizes both accuracy and FLOP. EfficientNet-B7 achieves state-of-the-art 84.3% top-1 accuracy on ImageNet while keeping inference time faster than prior ConvNets methods. Xie et al. introduced a multi-branch architecture called ResNext [21]. The deep network inherited by ResNets [22] consists of repeating building blocks aggregating a set of transformations with the same topology. By conducting extensive experiments, the authors came up with the conclusion that increasing cardinality is a more effective way of gaining accuracy than going deeper or wider, especially when depth and width start to give diminishing returns for existing models. The ResNext outperforms ResNet-101/152 [22], ResNet-200 [23], Inception-v3 [24] and Inception-ResNet-101/152 [25] on the ImageNet [26].

## III. METHODOLOGY

The fundamental step to tackle Object Detection in Aerial Images is to collect a sufficient amount of data for training models, however in real life, especially in aerial images, objects are often distributed randomly, leading to hardship in data preparation steps. Therefore, humans unintentionally create many imbalanced datasets (Fig. 2) and passively bias the deep learning models.

Object detection in aerial images appears challenging when detectors have to deal with the variety in object scales and orientations, making them extremely difficult to identify. The most common way to approach these problems is image augmentation [27] (more image-more feature-high performance). In addition, the Elhagry and Saeed proposed many methods in [28] to solve this problem, such as modifying generated anchor sizes for region proposals and investigating multiple backbones and loss functions and achieved an improvement of 4.7 mAP over the baseline.

Consider image augmentation as a feasible solution for these problems. Some basic data augmentation methods are applied frequently, such as random crop, random rotate, random flip, zoom, etc. These augmentation techniques only apply to the whole image (Fig. 3), so what about class imbalance? It seems extreme to improve classification performance via improving the classification branch inside the models due to common batch sampling methods.

What if we train the classification independently from the object detection network and ensemble the results together? Although this approach might look heavy, ensure that not only you can modify each class independently (augmentation, removing noises) but also add extra necessary data for extended training cycles (Fig. 4).

## IV. OUR APPROACH

### A. Pipeline

Our pipeline consists of two main parts: training and testing. In the training phase, Fig. 5, we train independently two networks (classification network and oriented object detection network). For the classification training data, we crop out oriented bounding boxes using convex-hull transformation and then carefully interpolate them to horizontal images. The data then proceeded to the classification network with a re-weighting mechanism. For the oriented object detection network, we train it with the original dataset to ensure regressed

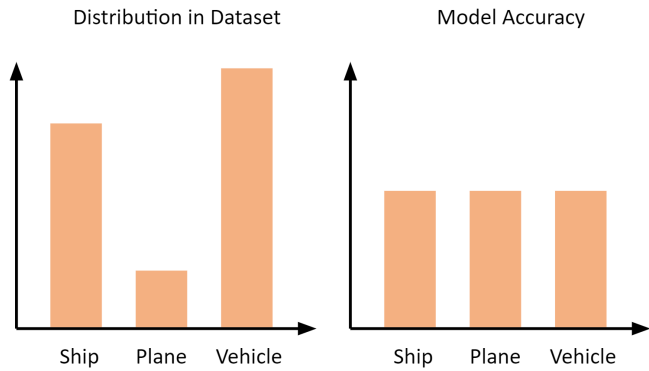


Fig. 2. Data Distribution Leads to the Problem of Class Imbalance in our Dataset. Our Target is to Implement the Model so that it Behaves Unbiased Toward Every Class.

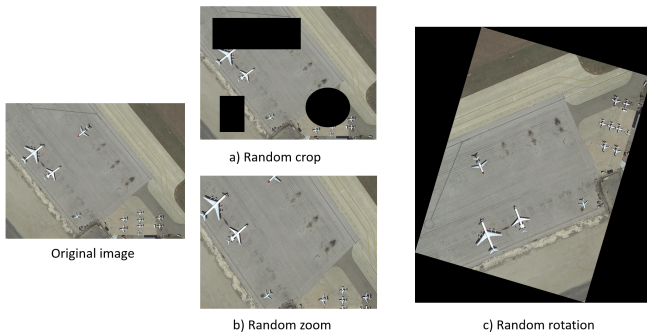


Fig. 3. Data Augmentation is a Common Method to Enrich the Amount of Data used for Training Object Detection Models. It only Enriches the Amount of Data, While we need the Instances Enrichment in Each Class.

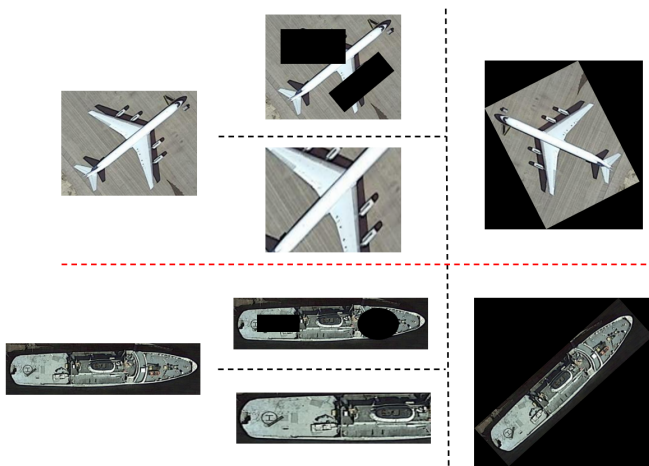


Fig. 4. Data Augmentation for Every Instance for each Class Ensures Richness in Features for the Classification Training Procedure.

bounding boxes are accurate. In the testing phase, Fig. 6, predicted oriented bounding boxes then be cropped out and interpolated to horizontal boxes. They will be re-labeled by the classification network (keeping high confidence score) and ensemble with the results of the oriented object detection network. Finally, those with low confidence scores will be removed as long with the NMS process from the prediction set.

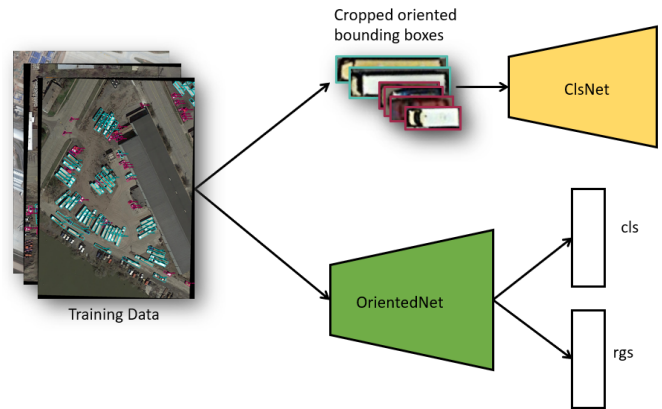


Fig. 5. Training Phase: Training Data Passed through an Oriented Object Detector's Network Contributes to Training Classification and Localization as in Other Well-Known Pipelines. However, the Required Data Preparation for Training the Classification Network is an Indispensable Step. Oriented Ground-Truths First are Interpolated to Cropped Oriented Bounding Boxes via Convex-Hull Transformation, then they're Transformed to Horizontal Images Fed to a Classification Network Together with not only the Re-Weighting Mechanism but also Image Augmentation Methods.

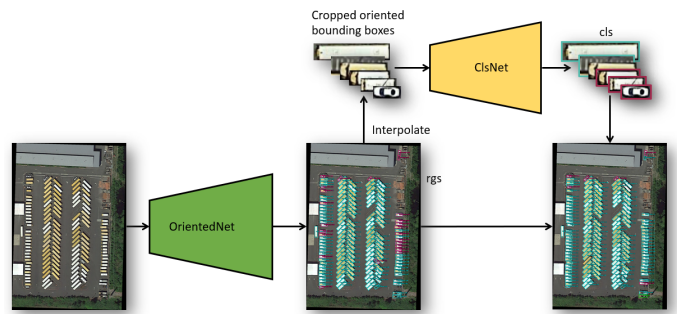


Fig. 6. Testing Phase: Different from the Training Phase, the Testing Image is Fed into Only the Oriented Object Detection Network. Then these Regressed Oriented Bounding Boxes Produced from the Network are Fed into the Classification Network after being Interpolated to Output the Predicted Label. Finally, we Ensemble the Outputs from Both Networks to Get the Appropriate Results in the Final Prediction set via Score Fine-Tuning and Non-Maximum Suppression.

### B. EfficientNet

EfficientNet [20] adopting the idea of the CNN network can be scaled in three dimensions: depth, width, and resolution. The depth of the neural network corresponds to the number of layers in the network. The width refers to the number of neurons in each layer or the number of channels in each Conv layer (the number of channels of the output). Resolution is simply the height and width of the input image. The following Equation 1 describes the compound scaling method where  $d$  is

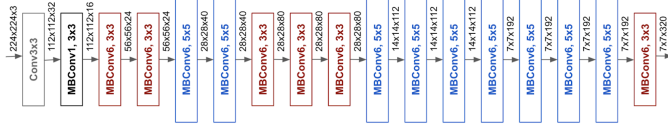


Fig. 7. EfficientNet-B0 [31] Structure.

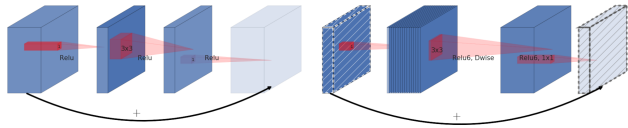


Fig. 8. Difference between Residual Block and Inverted Residual Block [30].

the depth,  $w$  is the width and  $r$  is the resolution. Moreover,  $\phi$  is a user-defined coefficient determining the available resources for model scaling and  $\alpha, \beta, \gamma$  specify how to assign these extra resources to network width, depth, and resolution.

$$\begin{aligned} d &= \alpha^\phi \\ w &= \beta^\phi \\ r &= \gamma^\phi \end{aligned} \tag{1}$$

Following the constraints:

$$\begin{aligned} \alpha \cdot \beta^2 \cdot \gamma^2 &\approx 2 \\ \alpha \geq 1, \beta \geq 1, \gamma \geq 1 \end{aligned}$$

The compound scaling method can generalize to existing CNN architectures such as MobileNet [29] and ResNet [22]. However, choosing the base network is important to get the best results because it only increases the predictive power of the neural network by reconstructing the parameters and structure of the base network. The author also uses Neural Architecture Search to build an efficient network architecture - EfficientNet B0. It achieves 77.3% accuracy on the ImageNet with 5.3M parameters and 0.39B FLOPS (ResNet-50 achieves 76% accuracy with 26M parameters and 4.1B FLOPS).

The main building block of EfficientNet-B0 is the MBConv block. The MBConv block, Fig. 8 is similar to the inverted residual block used in MobileNetv2 [30]. In this block, there is a shortcut connection between the beginning and the end of the block. The input is scaled with a 1x1 Conv layer to increase the number of channels or the depth of the feature map. Then they use Depthwise convolution 3x3 and Pointwise convolution (Conv layer 1x1) to reduce the number of channels of the output. A shortcut connection connects narrow layers (a small number of channels) while wider layers are in the middle of the shortcut connection (Fig. 7). This structure helps to reduce the number of parameters and the number of operations.

### C. R3Det

ReDet [4] proposed to take a huge step from horizontal object detection to oriented object detection by solving feature misalignment during the feature extraction process. The model uses the Feature Refinement Module together with feature interpolation to extract position information related to the

refined bounding box and reconstructs feature maps to achieve feature alignment. R3Det adopts Refined Rotation RetinaNet as the backbone with multiple stages of refinement like Cascade while speeding up the model by reducing the number of refined bounding boxes in the first stage.

### D. S2ANet

S2ANet [18] proposed to solve the inconsistency between classification and localization performance. It adopts the Feature Pyramid Network to extract high-level features. Anchor generator, namely Feature Alignment Module, generates high-quality anchors which then pass through Oriented Detection Module for classification and regression.

### E. RoI Transformer

RoI Transformer [16] tackles the problem of using horizontal bounding boxes (mismatch features). The model consists of the lightweight Rotated RoI learner with a 5 dimensions fully connected layer representing the offset of the rotated RoI corresponding to the HRoI. Rotated RoI warping generates fixed-size geometry robust features for classification and regression via feature maps and rotated RoIs.

### F. ReDet

ReDet [32], namely the Rotation-equivariant Detector adopts ResNet [22] and Feature Pyramid Network so as to extract rotation-equivariant features. RiRoI Align proposed along with the model transforms rotated RoIs generated by an RPN and RoI Transformer [16]. The final feature extraction step regresses final oriented bounding boxes and classifies corresponding labels.

### G. Oriented R-CNN

The well-designed two-stage detector Oriented R-CNN produces high-level features for oriented object detection. The main structure inherits from the two-stage object detector baseline while introducing a new representation of region proposals. The oriented RPN encodes features received from each level feature of the FPN [33] and decodes them into RoIs under the Midpoint Offset representation. Due to the Midpoint offset representation, the oriented proposal generated by oriented RPN is usually a parallelogram, so the model will slightly adjust the shorter diagonal to the same length as another diagonal obtaining oriented bounding boxes. Finally, each RoIs is divided into  $m \times m$  grids to produce a fixed-size feature map  $F'$ . The idea that produces feature map  $F'$  adopts the idea of the rotation transformation the same as [5] and these fixed-size feature maps  $F'$  are then fed into fully-connected layers regressing the offset and assigning categories for each oriented bounding box.

### H. Non Maximum Suppression

In the process of improving this result, after relabeling the result by EfficientNet, we use non-maximum suppression (NMS) to sift out the overlapping bounding box instances as follows (Fig. 9):

**Input:** a list of oriented bounding box  $A$ , corresponding confidence score  $S$  and overlap threshold  $N$ .



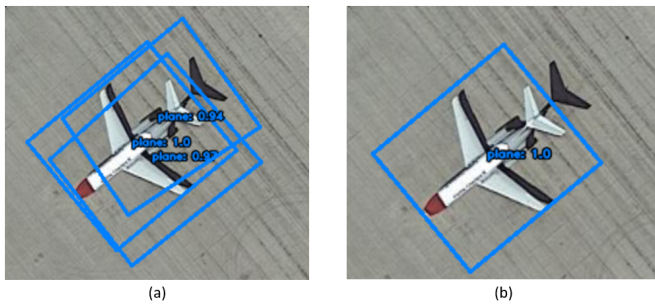


Fig. 9. Objectives of Non-Maximum Suppression. (a) Before NMS. (b) After NMS.

**Output:** a list of sifted bounding box B.

**Algorithm:**

Step 1: Select the bounding box with the highest confidence score, remove it from A and add it to the final list B.

Step 2: Now compare this bounding box with all the bounding boxes in A (calculate the IoU). If the IoU is greater than the threshold N, remove that bounding box from A.

Step 3: This process is repeated until there is no more bounding box left in A.

We calculate IoU between two oriented bounding boxes as IoU between two rectangles (each is the smallest horizontal rectangle that includes the corresponding oriented bounding box) as Fig. 10.

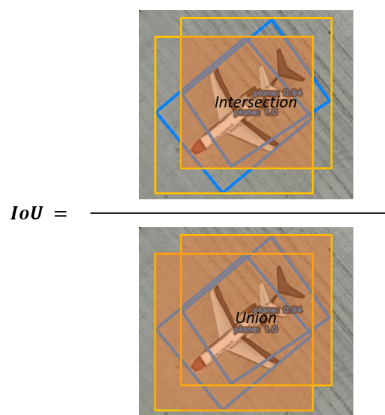


Fig. 10. Intersection over Union.

### I. Results Post-Processing

After having regressed the oriented bounding box, we transform it to the horizontal one and feed it through the classification network to get the appropriate label. Also we ensemble the classification results with the oriented object detection network to make cleaner results. Finally, NMS will be activated to sift and get the right fitter bounding boxes. This approach will ensure that there will be no more overlapping bounding boxes for one object caused by class imbalance and inconsistency between regression and classification (common problems in aerial images dataset).

## V. EXPERIMENT

### A. Experiment

**DOTA Oriented Dataset:** In this study, we conducted experiments on the DOTA dataset [34], which is a large-scale dataset widely used for the oriented object detection problem in aerial images. The DOTA was introduced in 2018, containing 2,806 images, and the proportion of the training set, validation set, and testing set were 1/2, 1/6, and 1/3, respectively. The dataset contains 188,282 instances which are accurately labeled of 15 common object categories includes: plane (PL), baseball-diamond (BD), bridge (BR), ground-track-field (GTF), small-vehicle (SV), large-vehicle (LV), ship (SH), tennis court (TC), basketball court (BC), storage tank (ST), soccer-ball-field (SBF), roundabout (RA), harbor (HA), swimming pool (SP), and helicopter (HC).

**Data Preparation for Classification Training:** To train the EfficientNet model, we need an input dataset that is an object image set (bounding box of objects) on each DOTA class. This process is illustrated in Fig. 11.

Step 1: Determine the object bounding box to be extracted [Fig. 11(a), red quadrangle].

Step 2: The object bounding box shape we need to take out is a rectangle but DOTA's labeled bounding box is arbitrary quadrilateral, so we take the smallest oriented bounding box that covers DOTA's labeled bounding box (convex-hull transformation) [Fig. 11(b), green bounding box].

Step 3: Rotate the image so that the bounding box to be taken becomes a horizontal bounding box. To reduce the calculation cost, we define the smallest area of the image that includes the box to be extracted, then proceed to rotate this area [Fig. 11(c), 11(d)]

Step 4: Extract the object bounding box [Fig. 11(e)].

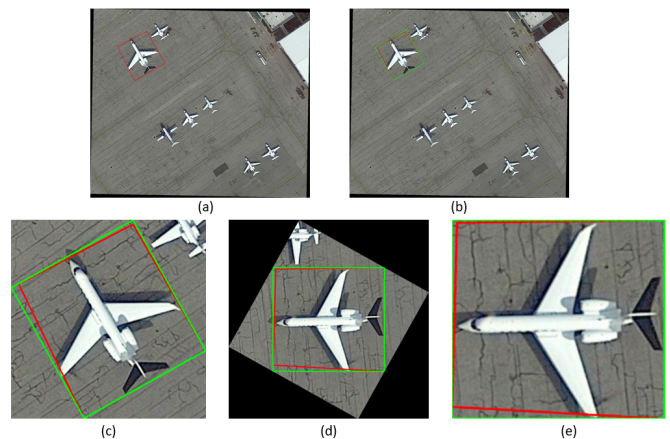


Fig. 11. Data Preparation Steps using Convex-Hull Transformation.

Apply the above steps to all instances on DOTA's train and validation set, then we get the train and validation set to train EfficientNet, the detail as shown in Fig. 12. DOTA is the imbalance dataset - a normal problem in real-life aerial images data where vehicles and ships outweigh most of them.

**Experimental Configuration:** We implement Oriented R-CNN, S2ANet [18], ReDet [32], R3Det [4], RoI Transformer

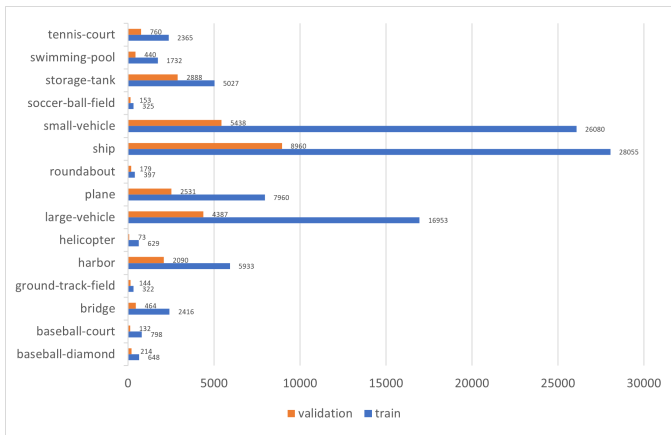


Fig. 12. Illustration of the Number of Ground-Truths in Training and Validation set of DOTA Dataset.

[16] on MMRotate [35] and EfficientNet-b3 on MMClassification [36] framework with repeat factor sampling methods [37] solving class imbalance problem using 2x GPU RTX 2080Ti

**Evaluation Metric:** To evaluate the performance of the models, and our pipeline on the DOTA dataset, we use the mAP score (mean Average Precision). Mean Average Precision is a popular evaluation metric used for object detection. It is the average of AP in every class in a dataset.

### B. Analysis

In this study, we test five models, R3Det, RoI Transformer, ReDet, S2Anet, and the Oriented R-CNN along with post-processing methods on the DOTA dataset and achieve the results presented in Table I.

For the initial results, all models perform very well in directional object detection (mAP ranges from 70% to 77%). In which the highest resulting model is ReDet (mAP = 76.68%) and the lowest is R3Det (mAP = 69.8%). Across DOTA's 15 layers, the accuracy of each model on each class varies widely (accuracy ranges from 50% to 90%), of which the highest are tennis-court (about 91%) and plane (about 89%); Especially the lowest is the helicopter class, resulting in a sizable difference between models (the highest is on ReDet 66.71%, the lowest is on R3Det 37.44%). Besides, R3Det also achieved very low results on the bridge class (45.41%) compared to other models in this class.

However, in the results of the above models, there is the same limitation which is the discovery of many other bounding boxes of the different classes located on the same instance (Fig. 13).

The two most accessible post-processing solutions:

- The first is the result of sifting out boxes, the scores of which are less than a certain threshold. The results are pretty good [Fig. 14(b)], but there are still some overlapping bounding boxes because these wrong bounding boxes have a score greater than the sifting threshold [Fig. 15(a)]. And this besides removing bounding boxes also removes quite a lot of true bounding boxes [Fig. 15(b)];

- The second is that as a result of using multi-class non-maximum suppression, similar to the first solution, the result of removing bounding boxes is also quite good [Fig. 14(c)]. However, this also has a limitation, which is the removal of bounding boxes in case 2 objects overlap. Both solutions caused the model's mAP results to decrease (about 2%).

After applying our initial solution: relabel all bounding boxes using EfficientNet, then apply non-maximum suppression to sift out overlapping bounding boxes. The result is that there is only one bounding box per object, which is more effective than all 2 conventional solutions above [Fig. 14(d)]. However, the accuracy of the results is significantly reduced (down almost 10-40% from the original). Among them, the most affected classes are soccer-ball-field (about 20-40%), bridge (about 14-60%), and helicopter (about 17-30%), ground-track-field (about 12-50%). The subjective reason for the decrease in results on layers is due to the misidentification between classes of EfficientNet, between classes with similar characteristics that make the model easily confused (between small-vehicle and large-vehicle, ship; between roundabout, plane and helicopter) (Fig. 16). The objective reason is that the imbalance of big data between classes (Fig. 12, small-vehicle and ship over 20k bounding boxes, while the remaining classes such as soccer-ball-field, ground-track-field are only about over 300 bounding boxes) makes the quality of layering between classes uneven; And the lack of information about the surroundings, because EfficientNet is only trained on the image, are bounding boxes (the similarity between ship and small-vehicle when cut into bounding boxes, Fig. 17).

So, to improve the accuracy of our pipeline, we combine the results between the original model and the results of EfficientNet, in detail: we will not be too confident in EfficientNet, which means that there will be no relabeling if the results returned by EfficientNet belong to soccer-ball-field, helicopter, bridge, and ship (which are classes that the EfficientNet model classifies inefficiently and affects the other classes analyzed above); it only be relabeled if the original model gives an uncertain result, which means that there will be a threshold if the model gives results below this threshold, then relabel will be applied.

The final score result of our pipeline is close to the original result (on three models with highest original results, RoI Transformer, ReDet, Oriented R-CNN), higher than the usual two solutions, and also very high efficiency in sifting wrong bounding boxes (Fig. 18). The other two models (S2Anet, R3Det) have a lot of wrong boxes, having a relabel doesn't work well either.

Besides efficiency, our pipeline is still wrong in several instances [Fig. 19(a)], which is still limited in cases where wrong bounding boxes or bounding boxes are part of the instance of the true bounding box (e.g. on container truck objects surrounded by large-vehicle boxes, but another bounding box covers the front of the car as a small-vehicle, Fig. 19(b))

In summary, our pipeline has solved the problem of multiple different boxes on the same object while keeping accuracy. The solution still does not resolve the case where wrong bounding boxes or different classes bounding boxes are part of the object of the true bounding box.



TABLE I. EXPERIMENTAL RESULTS

Model	Result Post-processing	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC	mAP
R3Det	Original	<b>89.3</b>	<b>75.31</b>	45.41	69.24	<b>75.54</b>	<b>72.89</b>	<b>79.29</b>	<b>90.89</b>	<b>81.02</b>	<b>83.26</b>	<b>58.82</b>	<b>63.15</b>	<b>63.41</b>	<b>62.21</b>	<b>37.41</b>	<b>69.8</b>
	Sifting	89.3	73.79	45.41	68.93	74.56	72.89	79.29	90.89	78.55	78.64	58.82	62.1	63.41	57.84	36.83	68.75
	Multi-class Non-maximum Suppression	89.3	75.31	<b>45.42</b>	<b>69.31</b>	74.32	72.67	79.28	90.89	79.06	82.71	55.23	62.71	63.33	62.1	33.25	69
	Relabel by EfficientNet	68.96	15.88	8.33	16.36	24.8	47.07	32.76	78.16	21.53	59.61	15.48	22.04	13.77	37.79	1.27	30.92
	Relabel by EfficientNet with conditions	84.4	72.48	44.6	53.2	53.31	64	69.53	85.29	77.15	60.12	57.38	61.84	56.7	41.46	31.36	60.86
S2ANet	Original	89.3	80.49	<b>50.42</b>	<b>73.23</b>	<b>78.42</b>	<b>77.4</b>	86.8	90.89	85.66	84.24	<b>62.16</b>	65.93	<b>66.66</b>	<b>67.76</b>	<b>53.56</b>	<b>74.19</b>
	Sifting	89.25	76.02	48.73	71.2	74.84	76.05	86.8	90.89	<b>85.99</b>	84.2	61.19	64.31	66.59	67.8	52.96	73.12
	Multi-class Non-maximum Suppression	<b>89.31</b>	<b>80.88</b>	50.42	71.56	76.49	76.16	<b>86.81</b>	<b>90.9</b>	85.23	<b>84.26</b>	59.49	<b>66.32</b>	66.66	67.76	53.56	74.19
	Relabel by EfficientNet	53.23	15.26	10.51	11.90	22.81	37.85	26.15	39.31	13.65	37.84	9.46	16.16	10.71	28.48	0.41	22.25
	Relabel by EfficientNet with conditions	82.96	71.9	40.05	64.93	37.11	61.47	75.67	16.32	76.05	12.35	4.55	64.4	58.65	13.42	43.12	48.2
RoI Transformer	Original	<b>88.98</b>	82.17	54.59	76.28	<b>79.29</b>	<b>77.96</b>	<b>87.94</b>	<b>90.91</b>	<b>87.19</b>	<b>85.65</b>	<b>61.44</b>	<b>62.63</b>	<b>74.63</b>	<b>72.43</b>	59.23	<b>76.09</b>
	Sifting	88.98	82.17	54.59	73.64	74.77	77.96	87.94	90.91	87.19	85.65	54.55	62.63	68.87	72.43	56.83	72.43
	Multi-class Non-maximum Suppression	88.98	<b>82.23</b>	<b>54.6</b>	<b>76.43</b>	74.77	77.96	87.94	90.91	86.76	85.54	60.83	62.63	74.57	72.43	56.86	75.78
	Relabel by EfficientNet	67.55	16.15	12.98	15.5	24.83	44.75	32.97	70.33	19.33	60.52	14.08	21.76	13.78	39.02	1.49	30.33
	Relabel by EfficientNet with conditions	88.85	82.22	54.59	76.31	78.64	77.83	87.9	90.9	87.12	85.55	61.05	62.62	68.84	72.39	<b>59.27</b>	75.6
ReDet	Original	89.2	83.77	52.21	71.04	78.05	<b>82.5</b>	<b>88.24</b>	<b>90.86</b>	<b>87.26</b>	<b>85.98</b>	<b>65.58</b>	62.86	<b>75.86</b>	70.04	66.71	<b>76.68</b>
	Sifting	<b>89.76</b>	78.79	47.01	65.2	<b>80.98</b>	80	87.33	90.74	79.17	86.23	49.09	<b>65.87</b>	65.75	<b>71.86</b>	55.21	72.87
	Multi-class Non-maximum Suppression	89.2	83.8	52.21	71.1	73.88	78.01	88.24	90.86	87.26	85.97	65.58	60.42	75.83	70.04	64.29	75.78
	Relabel by EfficientNet	68.81	16.11	8.36	16.38	24.82	46.27	32.76	70.13	21.77	59.74	15.55	22	13.72	37.82	1.3	30.37
	Relabel by EfficientNet with conditions	89.17	<b>83.81</b>	<b>52.21</b>	<b>71.1</b>	73.75	81.72	88.21	90.83	87.23	85.93	65.39	60.41	75.35	70	<b>66.89</b>	76.13
Oriented R-CNN	Original	89.35	81.41	<b>52.6</b>	<b>75.02</b>	<b>79.03</b>	<b>82.41</b>	87.82	90.9	86.4	<b>85.3</b>	<b>63.36</b>	<b>65.7</b>	<b>68.28</b>	<b>70.48</b>	<b>57.23</b>	<b>75.69</b>
	Sifting	89.35	81.41	52.6	72.58	74.3	77.96	87.82	90.9	86.4	85.3	63.36	63.68	68.28	70.48	54.53	74.6
	Multi-class Non-maximum Suppression	89.36	<b>81.45</b>	52.59	72.65	74.27	77.96	87.82	90.9	<b>86.57</b>	85.29	60.73	63.68	68.28	70.48	54.78	74.45
	Relabel by EfficientNet	89.18	74.79	38.19	60.78	71.14	67.22	77.21	90.87	77.35	78.19	43.58	56.21	63.9	61.6	37.06	65.82
	Relabel by EfficientNet with conditions	<b>89.36</b>	79.77	52.56	74.37	78.6	82.15	<b>87.82</b>	<b>90.9</b>	86.14	85.24	62.03	63.48	68.27	70.45	54.33	75.03

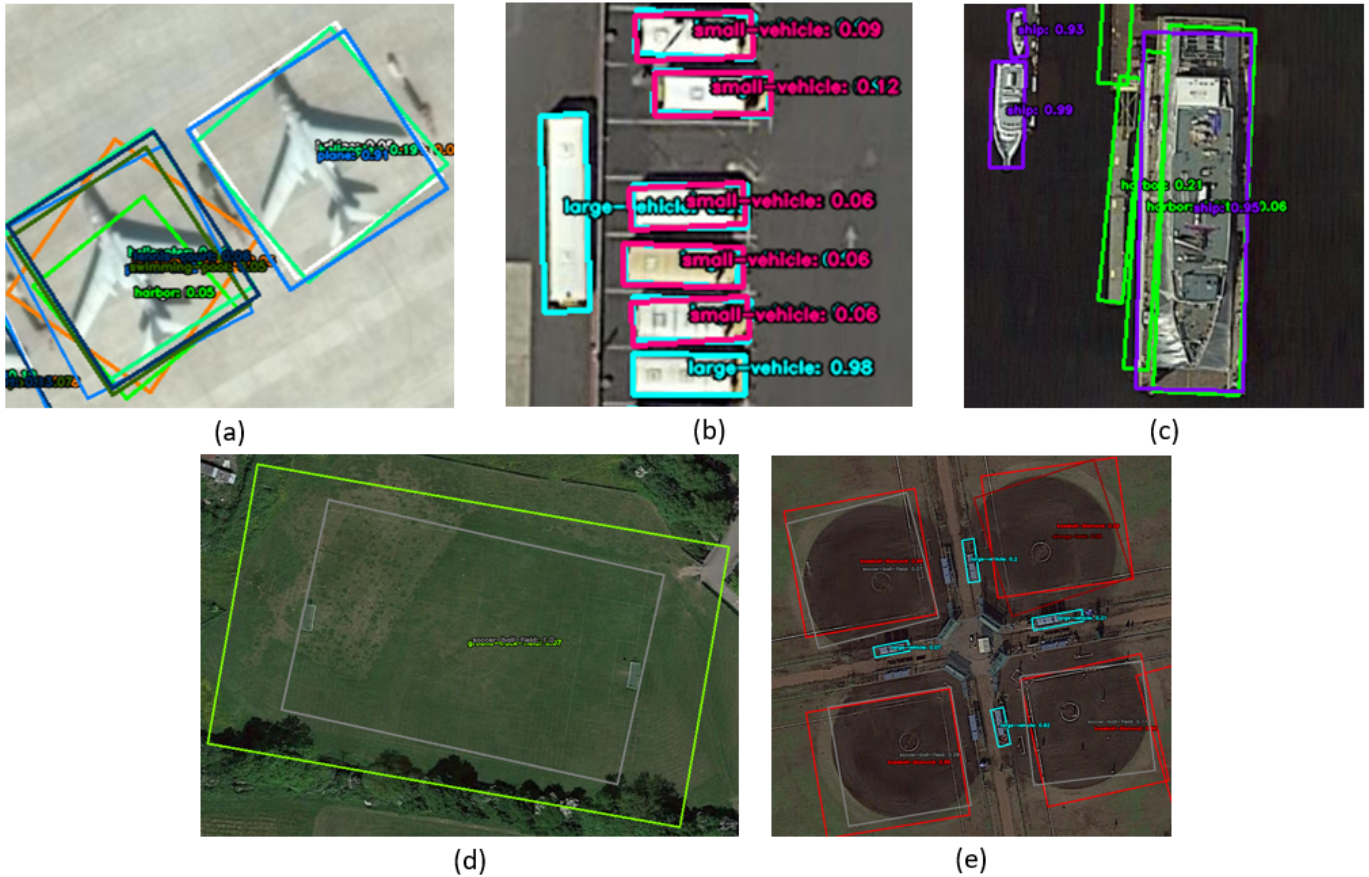


Fig. 13. Model's Result Limitation. (a) R3Det. (b) RoI Transformer. (c) ReDet. (d) S2ANet. (e) Oriented R-CNN

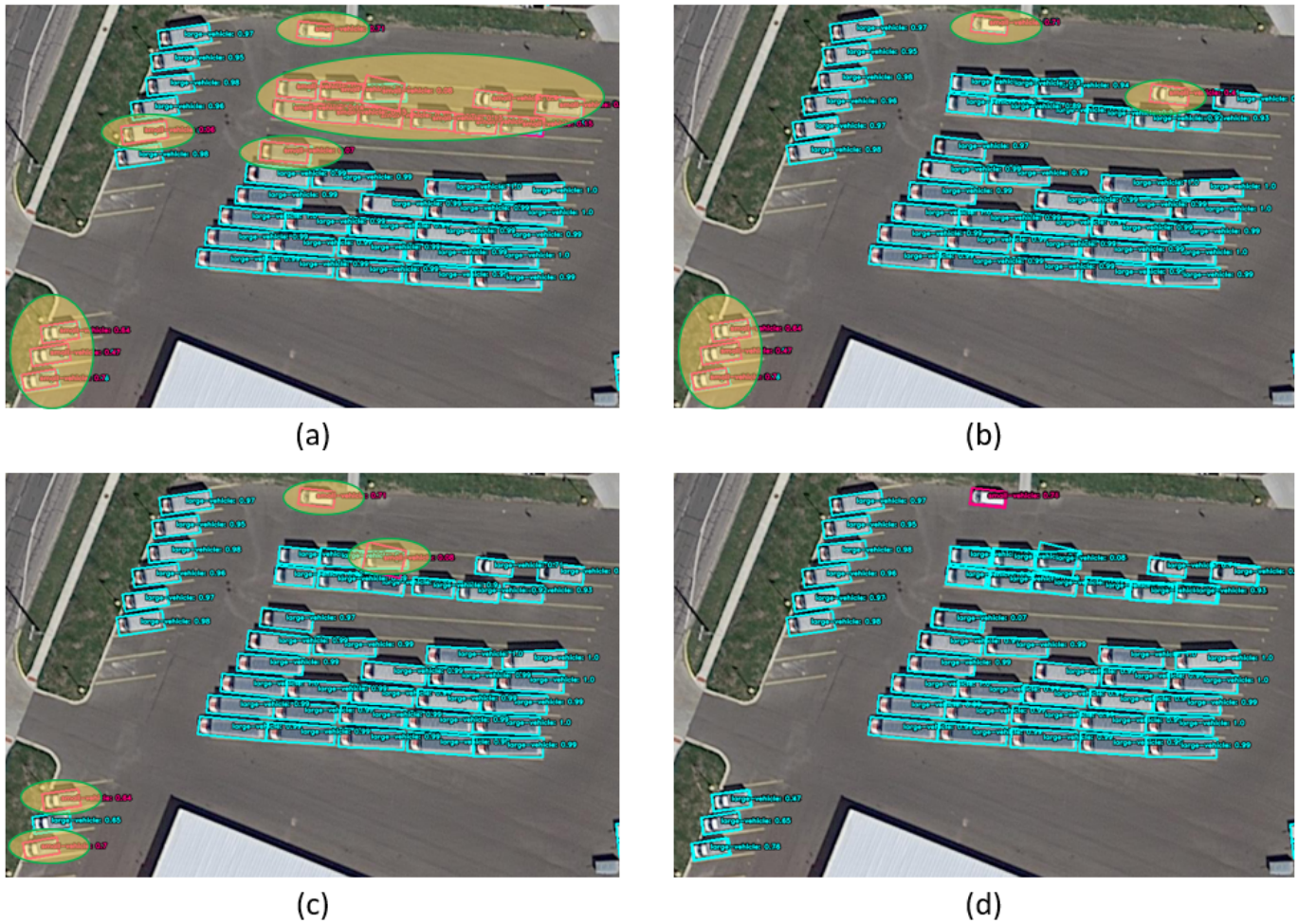


Fig. 14. Post-Processing Results. (a) Original. (b) Sifting. (c) Multi-Class Non-Maximum Suppression. (d) Relabel by EfficientNet. Covered Areas Represent Class Imbalance and Inconsistencies between Classification and Localization Leading to Bounding Boxes of Different Classes on One Object.

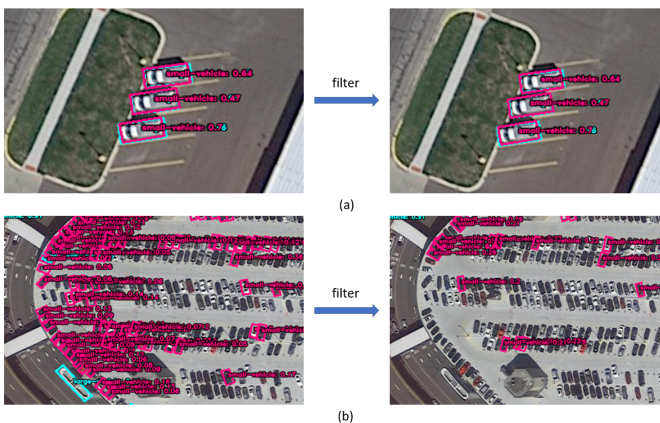


Fig. 15. Bad Case of Sifting Solution. (a) Wrong Bounding Boxes have a Score Greater than the Sifting Threshold. (b) True Bounding Boxes have a Score Smaller than the Sifting Threshold.

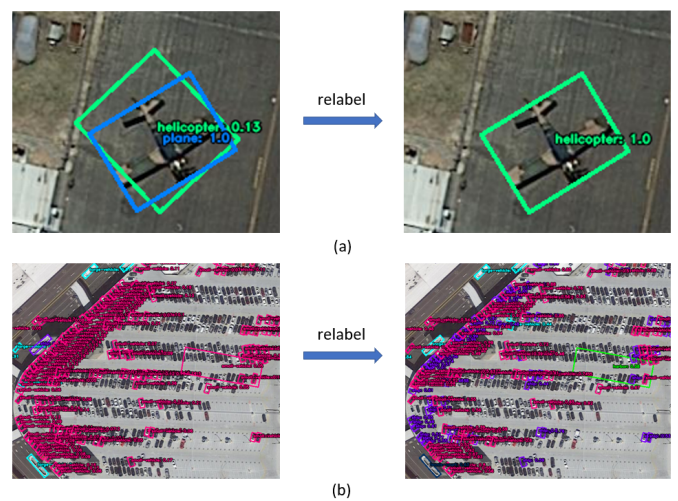


Fig. 16. Bad Case of Relabeling by EfficientNet. Misidentification between Classes of EfficientNet Resulting in the False Removal and Relabel Bounding Boxes



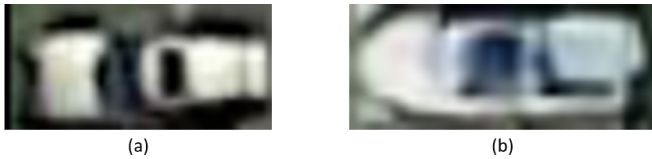


Fig. 17. Similarity between Small-Vehicle and Ship. (a) Small-Vehicle. (b) Ship.

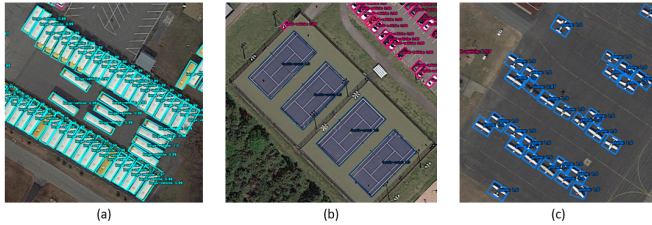


Fig. 18. Our Pipeline Result.

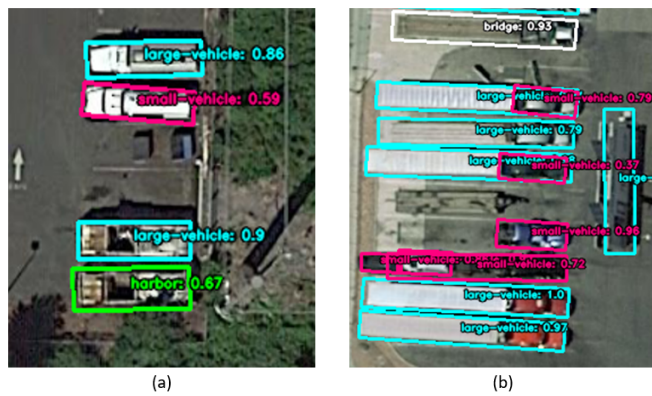


Fig. 19. Bad Case of our Pipeline. (a) Wrong Labeling. (b) Wrong Bounding Boxes of are Part of the Instance.

## VI. CONCLUSION

Generally, our pipeline adopts very well to many lots of SOTA baselines, yielding promising results and solving problems of inconsistency between classification and localization. According to our experimental results, our pipeline yields promising results on the oriented DOTA dataset by extracting oriented bounding boxes and feeding to independent training cycles. In the future, we are researching more training and testing pipelines, seeking more baselines for oriented object detection. Our work introduces a fine pipeline for tackling mismatched features in classification, if exploited well enough, it will significantly boost detection performance.

## ACKNOWLEDGMENT

This research is funded by Vietnam National University Ho Chi Minh City (VNU-HCM) under grant number DS2021-26-01. We also would like to show our gratitude to the UIT-Together research group for sharing their pearls of wisdom with us during this research.

## REFERENCES

- [1] S. N. K. B. Amit and Y. Aoki, "Disaster detection from aerial imagery with convolutional neural network," in *2017 International Electronics Symposium on Knowledge Creation and Intelligent Computing (IES-KCIC)*, 2017, pp. 239–245.
- [2] A. Ammar, A. Koubaa, M. Ahmed, A. Saad, and B. Benjdira, "Vehicle detection from aerial images using deep learning: A comparative study," *Electronics*, vol. 10, p. 820, 03 2021.
- [3] K. Nguyen, P. Nguyen, D. C. Bui, M. Tran, and N. D. Vo, "Analysis of the influence of de-hazing methods on vehicle detection in aerial images," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 6, 2022. [Online]. Available: <http://dx.doi.org/10.14569/IJACSA.2022.01306100>
- [4] X. Yang, J. Yan, Z. Feng, and T. He, "R3det: Refined single-stage detector with feature refinement for rotating object," 2019. [Online]. Available: <https://arxiv.org/abs/1908.05612>
- [5] J. Ding, N. Xue, Y. Long, G.-S. Xia, and Q. Lu, "Learning roi transformer for detecting oriented objects in aerial images," 2018. [Online]. Available: <https://arxiv.org/abs/1812.00155>
- [6] J. Ma, W. Shao, H. Ye, L. Wang, H. Wang, Y. Zheng, and X. Xue, "Arbitrary-oriented scene text detection via rotation proposals," *IEEE Transactions on Multimedia*, vol. 20, no. 11, pp. 3111–3122, nov 2018. [Online]. Available: <https://doi.org/10.1109/2Tmm.2018.2818020>
- [7] T. V. Le, H. N. N. Van, D. C. Bui, P. Vo, N. D. Vo, and K. Nguyen, "Empirical study of reppoints representation for object detection in aerial images," in *2022 IEEE Ninth International Conference on Communications and Electronics (ICCE)*, 2022, pp. 337–342.
- [8] A. Ahmad, H. Sakidin, M. Y. A. Sari, S. F. Sufahani *et al.*, "Naïve bayes classification of high-resolution aerial imagery," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 11, 2021.
- [9] R. Girshick, "Fast r-cnn," 2015. [Online]. Available: <https://arxiv.org/abs/1504.08083>
- [10] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," 2015. [Online]. Available: <https://arxiv.org/abs/1506.01497>
- [11] H. Zhang, H. Chang, B. Ma, N. Wang, and X. Chen, "Dynamic r-cnn: Towards high quality object detection via dynamic training," 2020. [Online]. Available: <https://arxiv.org/abs/2004.06002>
- [12] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable detr: Deformable transformers for end-to-end object detection," 2020. [Online]. Available: <https://arxiv.org/abs/2010.04159>
- [13] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot MultiBox detector," in *Computer Vision – ECCV 2016*. Springer International Publishing, 2016, pp. 21–37. [Online]. Available: [https://doi.org/10.1007/2F978-3-319-46448-0\\_2](https://doi.org/10.1007/2F978-3-319-46448-0_2)
- [14] Q. Chen, Y. Wang, T. Yang, X. Zhang, J. Cheng, and J. Sun, "You only look one-level feature," 2021. [Online]. Available: <https://arxiv.org/abs/2103.09460>
- [15] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, "Yolox: Exceeding yolo series in 2021," 2021. [Online]. Available: <https://arxiv.org/abs/2107.08430>
- [16] J. Ding, N. Xue, Y. Long, G.-S. Xia, and Q. Lu, "Learning roi transformer for detecting oriented objects in aerial images," 2018. [Online]. Available: <https://arxiv.org/abs/1812.00155>
- [17] Y. Jiang, X. Zhu, X. Wang, S. Yang, W. Li, H. Wang, P. Fu, and Z. Luo, "R2cnn: Rotational region cnn for orientation robust scene text detection," 2017. [Online]. Available: <https://arxiv.org/abs/1706.09579>
- [18] J. Han, J. Ding, J. Li, and G.-S. Xia, "Align deep features for oriented object detection," 2020. [Online]. Available: <https://arxiv.org/abs/2008.09397>
- [19] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," 2020. [Online]. Available: <https://arxiv.org/abs/2010.11929>
- [20] M. Tan and Q. V. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," 2019. [Online]. Available: <https://arxiv.org/abs/1905.11946>
- [21] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," 2016. [Online]. Available: <https://arxiv.org/abs/1611.05431>

- [22] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015. [Online]. Available: <https://arxiv.org/abs/1512.03385>
- [23] —, "Identity mappings in deep residual networks," 2016. [Online]. Available: <https://arxiv.org/abs/1603.05027>
- [24] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," 2015. [Online]. Available: <https://arxiv.org/abs/1512.00567>
- [25] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," 2016. [Online]. Available: <https://arxiv.org/abs/1602.07261>
- [26] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "Imagenet large scale visual recognition challenge," 2014. [Online]. Available: <https://arxiv.org/abs/1409.0575>
- [27] Q. M. Chung, T. D. Le, T. V. Dang, N. D. Vo, T. V. Nguyen, and K. Nguyen, "Data augmentation analysis in vehicle detection from aerial videos," in *2020 RIVF International Conference on Computing and Communication Technologies (RIVF)*, 2020, pp. 1–3.
- [28] A. Elhagry and M. Saeed, "Investigating the challenges of class imbalance and scale variation in object detection in aerial images," 2022. [Online]. Available: <https://arxiv.org/abs/2202.02489>
- [29] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," 2017. [Online]. Available: <https://arxiv.org/abs/1704.04861>
- [30] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," 2018. [Online]. Available: <https://arxiv.org/abs/1801.04381>
- [31] S. S. E. Mingxing Tan and G. A. Quoc V. Le, Principal Scientist. (2019) Efficientnet: Improving accuracy and efficiency through automl and model scaling. [Online]. Available: <https://ai.googleblog.com/2019/05/efficientnet-improving-accuracy-and.html>
- [32] J. Han, J. Ding, N. Xue, and G.-S. Xia, "Redet: A rotation-equivariant detector for aerial object detection," 2021. [Online]. Available: <https://arxiv.org/abs/2103.07733>
- [33] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," 2016. [Online]. Available: <https://arxiv.org/abs/1612.03144>
- [34] J. Ding, N. Xue, G.-S. Xia, X. Bai, W. Yang, M. Yang, S. Belongie, J. Luo, M. Datcu, M. Pelillo, and L. Zhang, "Object detection in aerial images: A large-scale benchmark and challenges," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2021. [Online]. Available: <https://doi.org/10.1109/TPAMI.2021.3117983>
- [35] Y. Zhou, X. Yang, G. Zhang, J. Wang, Y. Liu, L. Hou, X. Jiang, X. Liu, J. Yan, C. Lyu, W. Zhang, and K. Chen, "Mmrotate: A rotated object detection benchmark using pytorch," *arXiv preprint arXiv:2204.13317*, 2022.
- [36] M. Contributors, "Openmmlab's image classification toolbox and benchmark," <https://github.com/open-mmlab/mmlclassification>, 2020.
- [37] A. Gupta, P. Dollár, and R. Girshick, "Lvis: A dataset for large vocabulary instance segmentation," 2019. [Online]. Available: <https://arxiv.org/abs/1908.03195>

# TextBrew: Automated Model Selection and Hyperparameter Optimization for Text Classification

Rushil Desai<sup>1</sup>, Aditya Shah<sup>2</sup>, Shourya Kothari<sup>3</sup>, Aishwarya Surve<sup>4</sup> and Dr. Narendra Shekokar<sup>5</sup>  
Department of Computer Engineering, Dwarkadas J. College of Engineering  
Mumbai, India

**Abstract**—In building a machine learning solution, algorithm selection and hyperparameter tuning is the most time-consuming task. Automated Machine Learning is a solution to fully automate the process of finding the best model for a given task without actually having to try various models. This paper introduces a new AutoML system, TextBrew, explicitly built for the NLP task of text classification. Our system provides an automated method for selecting transformer models, tuning hyperparameters, and combining the best models into one by ensembling. Keeping in mind that new state-of-the-art models are being constantly introduced, TextBrew has been designed to be highly flexible and thus can support additional models easily. In our work, we experiment with multiple transformer models, each with numerous different hyperparameter settings, and select the most robust models. These models are then trained on multiple datasets to obtain accuracy scores, which are then used to build the meta-dataset to train the meta-model. Since text classification datasets are not as abundant, our system generates synthetic data to augment the meta-dataset using CopulaGAN, a deep generative model. The meta-model is an ensemble of five models, which predicts the best candidate model with an accuracy of 78.75%. The final model returned to the user is an ensemble of all the best models that can be trained under the given time constraint. Experiments on various datasets and comparisons with existing systems demonstrate the effectiveness of our system.

**Keywords**—Automated machine learning; AutoML; NLP; transformer models; hyperparameter optimization; CopulaGAN; generative model; meta-learning

## I. INTRODUCTION

As more and more people recognize the actual value of data and try to get the most out of it, the demand for machine learning tools is increasing. But the complexity of the tasks involved in machine learning can be overwhelming for non-ML experts, and this is where automated machine learning comes into the picture. Non-ML experts can use AutoML for building ML projects. ML experts can also use it to perform repetitive tasks to save time and accelerate machine learning research by automating the development of models. Using AutoML, one can train high-performing models without worrying about hyperparameters, model architecture, or cross-validation strategies [1]. It aims to automate the model selection process, as shown in [2], [3], and its applications are increasing by the day, hand-in-hand with the growth of applications of machine learning.

Machines must use complex data processing and state-of-the-art machine learning algorithms to work with natural language. It is also essential to decide the particular machine learning method based on the dataset and the task. Selecting the correct method may become difficult as more

accurate methods are constantly being developed. Even after considering all these factors for choosing the algorithms and methods, getting the best result is not assured. AutoML for text classification (TextBrew) can address all these challenges.

Finding the best model and the best set of hyperparameters is tedious. This task requires a lot of time and resources since training a model on a dataset may take hours or even days, subject to the size of the dataset and the model being used, and then changing hyperparameters and training the dataset all over again becomes an even more tedious and patience-testing task. We aim to tackle this issue and give the user the best model for an NLP classification task.

This paper presents a system that returns the trained model predicted to be the best for it within a particular time limit when given a dataset as an input. Our system is such that it is effortless to add new state-of-the-art transformer models to the pre-existing pool of models for the system to consider those models. In this paper, we have performed all experiments on BERT [4], ALBERT [5], and XLNet [6] for text classification using multiple hyperparameter combinations for each. A major part of building our AutoML system is creating the meta-dataset, which is made by training and testing several deep learning models on multiple text classification datasets. This process has been made efficient using generative models to synthesize data instead of running models on many datasets. The results demonstrate the effectiveness of our system.

### A. Contribution and Organization of the Paper

To sum up, we aim to reduce the time spent by a data scientist on selecting a particular model and its hyperparameter values. As seen in the next section, a significant drawback of many AutoML systems is the inability to support additional, more complex models. TextBrew has been developed keeping this problem in mind. The contributions of our paper are as follows:

- 1) We show the effectiveness of using deep generative models to synthesize the training data instead of spending hours running models on a large number of datasets.
- 2) TextBrew is a flexible system that can accommodate the new models with the least number of changes possible as it is not built around any specific set of models. All that is needed for the inclusion of a new model is for it to be added to the model pipeline which is run while generating the meta-dataset.

This paper is structured as follows. In the next section, we review the related literature. Section III discusses the

methodology we have adopted. The results are described in Section IV, followed by the discussion in Section V. Sections VI and VII outline the conclusion and a plan for future steps, respectively.

## II. REVIEW OF LITERATURE

AutoML is a relatively new research topic and has made significant progress in recent years. Many surveys [7] summarize the works of other researchers on this topic. Most of these works focus on individual modules such as neural architecture search (NAS) or hyperparameter optimization (HPO) in classifying tabular data using classical algorithms. There is significantly less work done on AutoML for text classification, which is quite different from classifying tabular data because textual data is unstructured and requires more computationally intensive algorithms.

There are works published that compare the performance of various existing AutoML tools on text classification tasks. Blohm et al. [8] attempt to answer the question of whether AutoML can be effectively applied to text classification or not. They compare the performance of four AutoML tools on thirteen text classification datasets and document how they perform as opposed to human-engineered models. Their experiments show that AutoML systems perform better than humans on 4 out of 13 tasks, and this number will continue increasing as more sophisticated AutoML tools for NLP tasks are developed.

Some individual works have tackled the problem of AutoML for text classification. Madrid et al. [9] propose a method that automatically builds text classification pipelines based on the metadata obtained by running experiments using standard algorithms on 81 datasets. While it is effective, metadata obtained from 81 datasets could be insufficient for a machine learning model to learn and give optimal results.

Wong et al. [10] incorporate transfer learning to reduce the computational cost of Neural AutoML. Their system learns the hyperparameter choices common to multiple tasks and uses this to accelerate the network design for a new task. The results show a reduction in convergence time for text classification. The drawback of this system is that meta-overfitting has not been dealt with, which is an issue.

Gomez et al. [11] have used a hyper-heuristic approach by employing a genetic algorithm to evolve a population of meta-rules to decide the best model for the particular text classification dataset. This work considers only simple models: K-Nearest Neighbor, Logistic Regression, Support Vector Machine, and Multinomial Naive Bayes. While this approach is interesting, it is not ideal because an AutoML system for the task of text classification needs to be able to handle the high computational cost and complexity of state-of-the-art transformer models.

A similar issue is present in the system proposed by Feurer et al. [12]. They use Bayesian optimization to capture the relationship between the hyperparameter values and the model performance. This system includes 15 classifiers, 14 feature preprocessing methods, and 4 data preprocessing method in its search space. Including deep learning methods would have made their system more competitive with other AutoML systems for text classification.

Additionally, there exist some open-source AutoML systems that work with text data. Jin et al. [13] propose the AutoML system AutoKeras, which efficiently performs neural architecture search using Bayesian optimization, enabling it to select the most profitable operation each time. Shi, Mueller, et al. [14] in their paper use transformer networks and stack ensemble them with classical tabular models to handle data that contain text, numeric, and categorical features. A drawback of this system is that it works with a limited search space compared to other AutoML tools.

We propose TextBrew, a simpler system for AutoML for text classification, which offers an automated way for selecting the best transformer models and then training these models on the dataset with hyperparameter tuning within the time constraints provided by the user. The issue of creating a large meta-dataset, which takes a tremendous amount of time and effort, is solved by using GANs [15] to synthesize additional data.

## III. METHODOLOGY

### A. Overview

The system offers an automated way for selecting the best transformer models and then training these models on the dataset with the proper hyperparameter settings. It is designed in such a manner that it can handle both binary as well as multiclass prediction datasets without explicitly stating so. As shown in Fig. 1, this system for automated text classification can be segregated into two parts:

- 1) Building and Training the Meta-model.
- 2) Using TextBrew for Model Prediction.

Meta-model refers to the model produced on training the meta-learning algorithm. In the first part, “Building and Training the Meta-model”, top-performing transformer models with many different combinations of hyperparameters are considered. After conducting experiments, the best-performing candidate models are shortlisted. “Candidate model” is used in this paper to refer to a model with a particular combination of hyperparameter values. Next, the multinomial Naive Bayes model is trained along with the selected candidate models on a collection of text classification datasets. Due to the small size of the meta-dataset (the dataset on which the meta-model is trained) generated at this stage, CopulaGAN [16] (a deep generative model) is used to synthesize additional data, helping our meta-model train better. From the meta-dataset, we use the accuracy of the Naive Bayes and ALBERT\_2 models along with the dataset size as the predictor variables to predict the performance of other models. A soft voting ensemble consisting of five models was used for building the meta-model, and this meta-model was trained on the generated dataset to achieve better accuracy.

In the second part, “Using TextBrew for Model Prediction”, the meta-model is used for model prediction. The best model predicted by the meta-model is trained on the dataset provided by the user under the given time constraint. After training the best candidate model predicted by the meta-model, the next best predicted model is trained if time remains. This is continued until the specified time is exceeded. Finally, an ensemble of all these models is returned to the user.



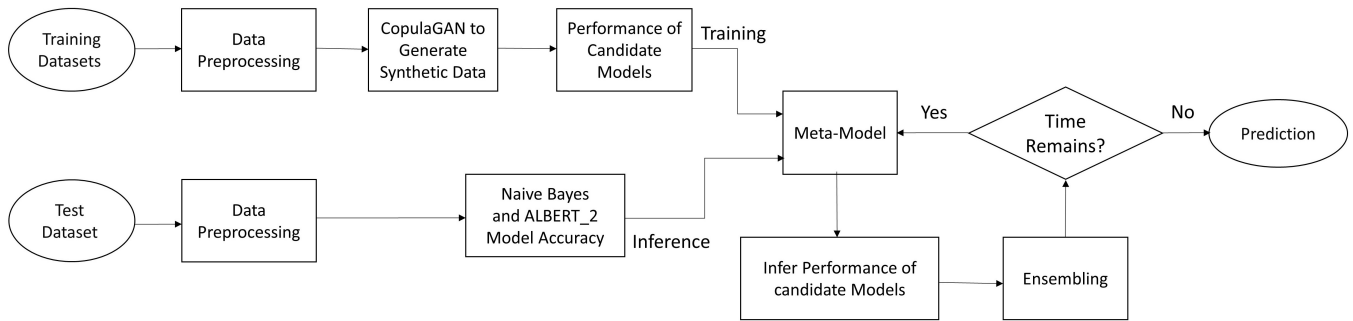


Fig. 1. Architecture Diagram of the TextBrew System.

### B. Building and Training the Meta-Model

Here, we describe how the meta-dataset is created and how the meta-model is trained.

For the real-world usage of this system, the system needs to be able to take into consideration a lot of different models. Transformer models are state-of-the-art machine learning models for NLP text classification tasks. There are a lot of different choices available, with GPT [17], BERT [4], RoBERTa [18], ELECTRA [19], ALBERT [5] and BART [20] being just a few of them. In fact, several different pre-trained models are based on each of these language models. For instance, BanglaBert [21] is a BERT-based Natural Language Understanding model pre-trained in Bangla. In addition, new models are being built at a blazing pace, and state-of-the-art models are being replaced frequently. This begs for a flexible system to accommodate the new models with the least number of changes possible.

1) *Candidate Model Selection:* Since transformer models are computationally costly, we perform all experiments needed to design this system using three models. The system is independent of the model choices and count. One can add more models in the same way described in this paper.

Gasparretto et al. [22], in their work, tested many algorithms on multiple text classification datasets, and it can be seen that these XLNet and BERT-base are among the top performing models in terms of accuracy.

In addition to these two models, we select ALBERT (A Light BERT) as one of our models as it has an architecture similar to that of BERT but has a fraction of the total number of parameters [5]. This makes it fast and less computationally expensive and so can help save time. Above are the reasons for selecting XLNet, BERT-base, and ALBERT for our experiments. These models are among the best for text classification.

Since all three transformer models have multiple hyperparameters that need to be tuned to an appropriate value, there are effectively many candidate models from which to choose. So, the next step is selecting the hyperparameter values for each model. We experimented with different combinations of hyperparameter values and picked the ones that gave us the best results when trained on multiple datasets.

The selected possible values of the hyperparameters for the models are shown in Table I. A short description of the

TABLE I. HYPERPARAMETERS AND SELECTED VALUES

Hyperparameters	Selected values
num_warmup_steps	[0, 1]
init_lr	[2e-5, 1e-5]
adam $\beta_1$	[0.8, 0.9]
adam $\beta_2$	[0.9, 0.999]
powers	[1, 1.5]
epochs	[3, 5]

hyperparameters is given below:

- 1) Num\_warmup\_steps: It is a parameter used to lower the learning rate to reduce the impact of deviating the model from learning on sudden new data set exposure.
- 2) Init\_lr: It is the initial learning rate before training and after warm-up steps.
- 3) Adam  $\beta_1$ , Adam  $\beta_2$ : The hyper-parameters  $\beta_1$  and  $\beta_2$  of Adam are initial decay rates used when estimating the moments of the gradient, which are multiplied by themselves at the end of each training step.
- 4) Powers: The power to use for polynomial decay.
- 5) Epochs: The number of passes or iterations of the training dataset by the machine learning model.

We choose the best set of hyperparameters by using an algorithm similar to grid-search. We run the transformer models with each possible combination of the hyperparameter values and compute the set of hyperparameters for which the model returns the best accuracy as output. In this case, each model effectively has  $2^6$  candidate models, counting all combinations of the six hyperparameters mentioned above.

Next, we selected three datasets to reduce bias in training the models. The three datasets are Sarcasm Detection [23], E-Mail classification NLP [24] and Financial phrase-bank [25]. Before training the candidate models on these datasets, the datasets were passed through a standard preprocessing pipeline. First, the text is converted to lowercase, and then we remove the URLs, non-ASCII characters, punctuations, and stopwords from the text. As this paper focuses mainly on model selection and hyperparameter optimization and not on the preprocessing of the data, the same preprocessing pipeline is used to clean all the datasets which are passed through the system.

All the candidate models for the three models, XLNet, BERT-base, and ALBERT, were then trained on the preprocessed datasets. Our system records how each hyperparameter combination performs on each dataset. Taking every possible combination of the hyperparameter values, we get 64 possible outcomes for each model. Hence, on trying these 64 outcomes for each of the three models for each of the three datasets, we effectively train 576 candidate models. The system records the accuracy attained by that particular model and the time required to train the model. From this dataset, we select the top six models (two from each of the three transformer models) as the shortlisted candidate models. We did this by averaging the accuracy of each unique hyperparameter combination model across the three datasets and taking the top two best-performing candidate models. This handles the situation where a model performs exceptionally well on one dataset and poorly on another. The top 6 models and the corresponding hyperparameter values are mentioned in Table II.

TABLE II. TOP SIX MODELS AND THEIR CORRESPONDING HYPERPARAMETERS

Model Name	Hyperparameters				
	epochs	warmup_steps	learning_rate	adam $\beta_1$	adam $\beta_2$
BERT_1	5	0	2e-5	0.9	0.999
BERT_2	5	0	1e-5	0.9	0.9
ALBERT_1	5	1	2e-5	0.8	0.9
ALBERT_2	3	0	1e-5	0.9	0.9
XLNet_1	3	1	1e-5	0.8	0.9
XLNet_2	3	1	2e-5	0.8	0.999

2) *Meta-Dataset Preparation:* After selecting the top six models, we create the meta-dataset for training the meta-model. We curated a list of 44 datasets consisting of both binary and multiclass text classification datasets. A longer list of datasets would have been beneficial, but for the task of English language text classification, the datasets publicly available are limited in number. On the other hand, a longer list of datasets means a lot more time to build the meta-dataset as more model training would have to be done. We solve this problem by synthesizing data using GANs, as described later in the paper. For each dataset, our system automatically trains all the selected candidate models and one additional multinomial Naive Bayes model. The accuracy and the training time taken for all these models are recorded.

The pipeline for training the multinomial Naive Bayes model consists of a count vectorizer that transforms the text into a vector based on the frequency of each word. A count matrix is created in which a column of the matrix represents each unique word in the dataset, and each row of the dataset is a row in the matrix, and the cells contain the count of the word in that text. Then, we transform the count matrix into a normalized TF-IDF representation. We use TF-IDF instead of the raw frequencies to reduce the impact of the frequently occurring words that are less informative than other words that occur in a small fraction of the dataset. We then train the multinomial Naive Bayes classifier on this. The reason for considering Naive Bayes for our model prediction is stated in the next section.

Table III shows a few records from our final meta-dataset.

3) *Feature Extraction:* The prerequisite step for training a classifier is feature extraction. Text embedding is the most

popular feature extraction method for text-oriented tasks [29]. Calculating embeddings for entire datasets is not feasible because of the computational expense that comes with it. Since the performance of any model depends on the dataset it has been trained on, the performance metrics of a model can be treated as a feature that is representative of the dataset. This approach is the basis for feature extraction in this paper.

Above is the reason for including multinomial Naive Bayes in the list of models. The accuracy of the Naive Bayes model is part of the feature space. The reason for choosing Naive Bayes is that it is extremely fast to train on the datasets compared to other models, making it efficient for use as a feature extraction tool. We also use the number of records in the dataset to predict the accuracy values of various models. Further, since ALBERT is A Light BERT model, it is faster to train than other transformer models we have chosen. Studying Table IV, we see that ALBERT\_2 consistently takes the least time to train. This makes it possible for us to take ALBERT\_2 accuracy scores as the third predictor variable to expand the feature space and, in turn, increase the accuracy of our meta-model. After running experiments to validate our hypothesis, we see that using the performance of the ALBERT\_2 model along with Naive Bayes performance and the dataset size improves the accuracy of the meta-model for predicting the best model. Adding the ALBERT\_2 accuracy score to the feature space boosts the accuracy by 26%, from 0.625 to 0.788, as shown in Fig. 2.

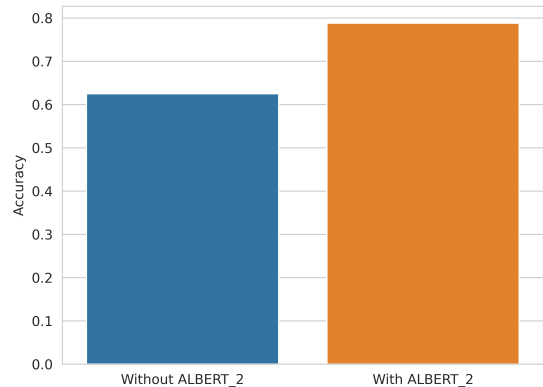


Fig. 2. Improvement in Overall Accuracy after Adding ALBERT\_2 Accuracy Score to the Feature Space.

4) *Using Generative Models to Synthesize Data:* Creating an extensive dataset requires a lot of time, and even if time constraints were not to be considered, publicly available text classification datasets are limited in number. Furthermore, overfitting is a problem faced by a model when the model predicts exceedingly well on the training dataset but cannot provide similar results on the testing dataset. It is known that small datasets are prone to overfitting. As a result, we need to tackle the overfitting problem caused by the small size of the dataset.

In this paper, we use GANs [15] to synthesize data and increase the dataset size to obtain a better result from the meta-classifier. Generative adversarial networks are an unsupervised learning approach to generative modeling using deep learning methods such as convolutional neural networks (CNN). Generative modeling involves automatic learning of patterns and the

TABLE III. CANDIDATE MODELS AND THEIR ACCURACIES ON DIFFERENT DATASETS

Dataset	Naive Bayes	BERT_1	BERT_2	ALBERT_1	ALBERT_2	XLNet_1	XLNet_2	Best Model
Cyberbullying [26]	0.8885	0.9925	0.9925	0.9950	0.9925	0.9925	0.9943	ALBERT_1
WELFake [27]	0.8510	0.9657	0.9531	0.9743	0.9788	0.9748	0.9834	XLNet_2
Amazon [28]	0.8550	0.9343	0.9086	0.9571	0.8200	0.6229	0.7257	ALBERT_1
Yelp [28]	0.7800	0.9486	0.8571	0.9371	0.8286	0.6514	0.7743	BERT_1
IMDb [28]	0.8000	0.9313	0.8053	0.9504	0.7939	0.6603	0.8397	ALBERT_1

TABLE IV. CANDIDATE MODELS AND THEIR TRAINING TIME (IN SECONDS) ON DIFFERENT DATASETS

Dataset	Naive Bayes	BERT_1	BERT_2	ALBERT_1	ALBERT_2	XLNet_1	XLNet_2	Fastest Model
Cyberbullying [26]	0.16	479.21	478.63	496.30	298.16	431.29	431.49	ALBERT_2
WELFake [27]	1.46	344.80	330.15	336.70	202.68	296.60	287.91	ALBERT_2
Amazon [28]	0.02	82.80	65.37	66.66	39.45	68.52	57.11	ALBERT_2
Yelp [28]	0.01	65.74	65.80	65.53	39.42	57.38	57.40	ALBERT_2
IMDb [28]	0.02	49.85	49.70	49.14	29.83	43.66	43.62	ALBERT_2

regularities of the data in a way that can be used to synthesize new examples that plausibly could have been drawn from the original dataset.

We employ the CopulaGAN model [16] to generate 150 data points for each group of datasets having a particular candidate model as the best model. This ensures that the resulting dataset is not unbalanced. The CopulaGAN model is a variation of the CTGAN model [30] which takes advantage of the CDF-based transformation that the GaussianCopulas apply to make the underlying CTGAN model task of learning the data easier. We make use of different metrics to compare the synthesized data and the original data. Seeing the results in Table V, we can be assured of the quality of the synthetic data.

TABLE V. METRICS FOR SYNTHETIC DATA EVALUATION

Kolmogorov-Smirnov Test	Multiclass Decision Tree Classifier
0.69	0.59

The Kolmogorov-Smirnov test [31] is based on the maximum difference between an empirical and a hypothetical cumulative distribution. The test concerns the agreement between the generated data and the original data. The Multiclass Decision Tree Classifier test lets the generated data pass through a decision tree generated on the original dataset. The resulting accuracy of this test indicates the percentage of data points that fall perfectly in accordance with the original data.

5) *Meta-Model Details:* We use a soft voting ensemble as our meta-model. This means that we predict the class with the largest summed probability from all the models that are part of the ensemble. Our ensemble consists of five models, each of which can perform a multiclass classification.

- 1) Multinomial Logistic Regression [32]
- 2) XGBClassifier [33]
- 3) C-Support Vector Classification following the One-vs-Rest scheme [34]
- 4) Random Forest Classifier [35]
- 5) AdaBoost Classifier [36]

We then use the generated data to train the meta-model. To see how the synthesized data improves the meta-model

performance, we train it on the original data, record the accuracy, and then do the same on the synthetic data. As expected, we get a low accuracy of 0.475 on training the meta-model on a small dataset. This accuracy score is boosted to 0.7875 after using synthetic data, which consists of a lot more data.

### C. Using TextBrew for Model Prediction

In this section, we describe how the user would use our proposed system and how the system makes use of the meta-model we trained in the previous section. The user passes a text classification dataset and an optional parameter: *allowed\_training\_time* as input to the system. The parameter *allowed\_training\_time* denotes the upper limit for the time the user expects the trained model to be returned as an output. This parameter helps users set a time limit for training the model if they have a time constraint. The default value of this parameter is set as 15 minutes in our experiments.

First, the dataset is passed through the same preprocessing pipeline we created for the datasets used for training, which is mentioned in the above sections. The preprocessing pipeline helps to clean the data and eliminate inconsistencies. The next step is to train Naive Bayes and ALBERT\_2 models on the preprocessed dataset. We considered Naive Bayes because it is swift to train, as shown in Table IV. ALBERT\_2 was considered because it was the fastest to train out of the six transformer models and could be used as a feature to predict the best model. The output of these two models, plus the size of the dataset, forms the feature space. These features are input to the meta-model, which predicts the best-performing model. The dataset given as input to the system is then trained on this predicted model, and the time taken for training is monitored. After training is completed, if the system has not exceeded the *allowed\_training\_time*, then it trains the next best model predicted on the same dataset. This continues until it exceeds the allowed time. Finally, an ensemble of all these trained models is created using a soft voting ensemble algorithm. The prediction probabilities for all the class labels are summed up, and the class label with the most considerable sum is selected as the predicted class. This ensemble is returned as the output.

#### IV. RESULTS

As mentioned in the previous section, we developed a meta-model that correctly predicted the best model with an accuracy of 78.75% on using CopulaGAN to synthesize data, compared to 47.5% on using the original data. This is a strong score given that the meta-dataset is generated using only 44 datasets because of the limited availability of open-source datasets for text classification. Additionally, 95% of the time, the first model predicted is among the top three models in terms of accuracy. These numbers indicate that it is highly likely that the best model will be part of the ensemble returned by the system as it consists of the first few models predicted to perform well by the meta-model.

We also present the results of the experiments that were performed using different datasets and different existing AutoML solutions for text classification. FakeNews, Covid Tweets Sentiment Analysis, and Cyberbullying Classification are the datasets used. We compare our system with AutoKeras [13] and AutoGluon [14]. In order to create an unbiased environment for all tests, we use Intel® Xeon® 2.00GHz CPU, Tesla P100 16GB GPU, and 13 GB of free memory for our usage.

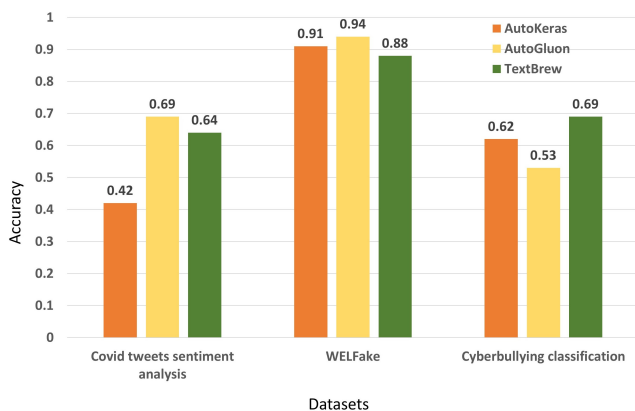


Fig. 3. Comparison of AutoML Systems, Including TextBrew (green), AutoKeras (orange) and AutoGluon (yellow).

We can see from Fig. 3 that our system gives the best accuracy on one dataset and a reasonable accuracy on the rest compared to AutoKeras and AutoGluon. Given that this is a prototype of our system and has only six candidate models originating from three different transformer models, i.e., BERT-base, XLNet, and ALBERT, the performance, as seen in Fig. 3, is extremely promising. Increasing the number of models described in the methodology section would lead to much better performance.

#### V. DISCUSSION

Today, very few AutoML systems for the NLP task of text classification exist that incorporate state-of-the-art deep learning models. AutoGluon is one of the best-performing AutoML models for text classification, as seen in Fig. 3, and TextBrew follows closely behind. In fact, TextBrew beats AutoGluon on one out of the three datasets with an accuracy that is 16% higher. With access to more computing power, we can incorporate more state-of-the-art deep learning models and a more comprehensive range of hyperparameter options, which

will enable TextBrew to be competitive and could challenge the best AutoML tools present today.

Moreover, TextBrew returns an ensemble of models that would give the best results on the given dataset. The best possible model out of the six selected models is found to be in the top three with an accuracy of 95%, of which 78.75% constitutes the situation where the first model predicted is actually the best. These numbers convey that the probability of the best candidate model being included in the ensemble generated by the system is very high.

The workings of our system back up the suggestion that automated machine learning does not have to be complicated to give good results. This is because this system does not employ any complicated methods or algorithms like genetic algorithms as seen in [11], deep reinforcement learning as seen in [10] and Bayesian optimization as visited in [12], [13]. The approach here is to build a meta-dataset and train classical ML models, which are used to predict the suitable transformer model. Nevertheless, TextBrew performs competitively considering the dataset's size and the approach's simplicity.

#### VI. CONCLUSION

This paper proposes TextBrew: an automated machine learning system for text classification. As discussed in the previous section, TextBrew suggests that AutoML systems do not have to be highly complex to perform well. The meta-model predicts the best possible model and suitable hyperparameters while considering the user's time constraint. Our meta-model predicts one of the top three models 95% of the time, with the best candidate model being predicted with an accuracy of 78.75%. This means that it is highly likely that the ensemble created by the system will include the best models.

The final model returned to the user is an ensemble of all the top candidate models that can be trained under the given time constraint. As seen in the results section, considering the low number of models in our system, TextBrew is promising, and it backs the idea that automated machine learning is not as complex as one would think.

#### VII. FUTURE WORK

For our future work, we aim to manually build and compile a greater number of text classification datasets to train our meta-model better. Additionally, we aim to access a more powerful computing system to include a much larger number of models and hyperparameter choices in our system.

#### REFERENCES

- [1] S. K. Karmaker ("Santu"), M. M. Hassan, M. J. Smith, L. Xu, C. Zhai, and K. Veeramachaneni, "Automl to date and beyond: Challenges and opportunities," *ACM Comput. Surv.*, vol. 54, no. 8, oct 2021. [Online]. Available: <https://doi.org/10.1145/3470918>
- [2] M. Feurer, A. Klein, K. Eggenberger, J. Springenberg, M. Blum, and F. Hutter, "Efficient and robust automated machine learning," in *Advances in Neural Information Processing Systems*, C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, Eds., vol. 28. Curran Associates, Inc., 2015. [Online]. Available: <https://proceedings.neurips.cc/paper/2015/file/11d0e6287202fced83f79975ec59a3a6-Paper.pdf>

- [3] N. Fusi, R. Sheth, and M. Elibol, "Probabilistic matrix factorization for automated machine learning," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, ser. NIPS'18. Red Hook, NY, USA: Curran Associates Inc., 2018, p. 3352–3361.
- [4] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," *CoRR*, vol. abs/1810.04805, 2018. [Online]. Available: <http://arxiv.org/abs/1810.04805>
- [5] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "ALBERT: A lite BERT for self-supervised learning of language representations," *CoRR*, vol. abs/1909.11942, 2019. [Online]. Available: <http://arxiv.org/abs/1909.11942>
- [6] Z. Yang, Z. Dai, Y. Yang, J. G. Carbonell, R. Salakhutdinov, and Q. V. Le, "Xlnet: Generalized autoregressive pretraining for language understanding," *CoRR*, vol. abs/1906.08237, 2019. [Online]. Available: <http://arxiv.org/abs/1906.08237>
- [7] X. He, K. Zhao, and X. Chu, "Automl: A survey of the state-of-the-art," *Knowledge-Based Systems*, vol. 212, p. 106622, 2021. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0950705120307516>
- [8] M. Blohm, M. Hanussek, and M. Kintz, "Leveraging automated machine learning for text classification: Evaluation of automl tools and comparison with human performance," pp. 1131–1136, 01 2021.
- [9] J. Madrid, H. J. Escalante, and E. Morales, "Meta-learning of textual representations," 2019. [Online]. Available: <https://arxiv.org/abs/1906.08934>
- [10] C. Wong, N. Hounsby, Y. Lu, and A. Gesmundo, "Transfer learning with neural automl," *Advances in neural information processing systems*, vol. 31, 2018.
- [11] J. C. Gomez, S. Hoskens, and M.-F. Moens, "Evolutionary learning of meta-rules for text classification," in *Proceedings of the Genetic and Evolutionary Computation Conference Companion*, ser. GECCO '17. New York, NY, USA: Association for Computing Machinery, 2017, p. 131–132. [Online]. Available: <https://doi.org/10.1145/3067695.3075601>
- [12] M. Feurer, A. Klein, K. Eggensperger, J. T. Springenberg, M. Blum, and F. Hutter, "Efficient and robust automated machine learning," in *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, ser. NIPS'15. Cambridge, MA, USA: MIT Press, 2015, p. 2755–2763.
- [13] H. Jin, Q. Song, and X. Hu, "Auto-keras: An efficient neural architecture search system," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 2019, pp. 1946–1956.
- [14] X. Shi, J. Mueller, N. Erickson, M. Li, and A. Smola, "Multimodal automl on structured tables with text fields," in *8th ICML Workshop on Automated Machine Learning (AutoML)*, 2021.
- [15] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, Eds., vol. 27. Curran Associates, Inc., 2014. [Online]. Available: <https://proceedings.neurips.cc/paper/2014/file/5ca3e9b122f61f8f06494c97b1afcf3-Paper.pdf>
- [16] N. Patki, R. Wedge, and K. Veeramachaneni, "The synthetic data vault," in *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, Oct 2016, pp. 399–410.
- [17] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever *et al.*, "Improving language understanding by generative pre-training," 2018.
- [18] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized BERT pretraining approach," *CoRR*, vol. abs/1907.11692, 2019. [Online]. Available: <http://arxiv.org/abs/1907.11692>
- [19] K. Clark, M. Luong, Q. V. Le, and C. D. Manning, "ELECTRA: pre-training text encoders as discriminators rather than generators," *CoRR*, vol. abs/2003.10555, 2020. [Online]. Available: <https://arxiv.org/abs/2003.10555>
- [20] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," *CoRR*, vol. abs/1910.13461, 2019. [Online]. Available: <http://arxiv.org/abs/1910.13461>
- [21] A. Bhattacharjee, T. Hasan, K. Samin, M. S. Rahman, A. Iqbal, and R. Shahriyar, "Banglabert: Combating embedding barrier for low-resource language understanding," *CoRR*, vol. abs/2101.00204, 2021. [Online]. Available: <https://arxiv.org/abs/2101.00204>
- [22] A. Gasparetto, M. Marcuzzo, A. Zangari, and A. Albarelli, "A survey on text classification algorithms: From text to predictions," *Information*, vol. 13, no. 2, p. 83, Feb 2022. [Online]. Available: <http://dx.doi.org/10.3390/info13020083>
- [23] K. Zeynalzade, "Sarcasm detection," Nov 2021. [Online]. Available: <https://www.kaggle.com/datasets/theynalzade/news-headlines-for-sarcasm-detection>
- [24] A. Miglani, "E-mail classification nlp," Sept 2020. [Online]. Available: <https://www.kaggle.com/datasets/datatattle/email-classification-nlp>
- [25] P. Malo, A. Sinha, P. Korhonen, J. Wallenius, and P. Takala, "Good debt or bad debt: Detecting semantic orientations in economic texts," *Journal of the Association for Information Science and Technology*, vol. 65, no. 4, pp. 782–796, 2014.
- [26] J. Wang, K. Fu, and C.-T. Lu, "Sosnet: A graph convolutional network approach to fine-grained cyberbullying detection," in *2020 IEEE International Conference on Big Data (Big Data)*. IEEE, 2020, pp. 1699–1708.
- [27] P. K. Verma, P. Agrawal, I. Amorim, and R. Prodan, "Welfare: Word embedding over linguistic features for fake news detection," *IEEE Transactions on Computational Social Systems*, vol. 8, no. 4, pp. 881–893, 2021.
- [28] D. Kotzias, M. Denil, N. de Freitas, and P. Smyth, "From group to individual labels using deep features," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '15. New York, NY, USA: Association for Computing Machinery, 2015, p. 597–606. [Online]. Available: <https://doi.org/10.1145/2783258.2783380>
- [29] H. Liang, X. Sun, Y. Sun, and Y. Gao, "Text feature extraction based on deep learning: a review," *EURASIP Journal on Wireless Communications and Networking*, vol. 2017, no. 1, p. 211, Dec 2017. [Online]. Available: <https://doi.org/10.1186/s13638-017-0993-1>
- [30] L. Xu, M. Skoularidou, A. Cuesta-Infante, and K. Veeramachaneni, "Modeling tabular data using conditional gan," 2019. [Online]. Available: <https://arxiv.org/abs/1907.00503>
- [31] F. J. Massey Jr, "The kolmogorov-smirnov test for goodness of fit," *Journal of the American statistical Association*, vol. 46, no. 253, pp. 68–78, 1951.
- [32] C. Kwak and A. Clayton-Matthews, "Multinomial logistic regression," *Nursing research*, vol. 51, no. 6, pp. 404–410, 2002.
- [33] T. Chen and C. Guestrin, "XGBoost," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, aug 2016. [Online]. Available: <https://doi.org/10.1145/2939672.2939785>
- [34] C.-C. Chang and C.-J. Lin, "Libsvm: A library for support vector machines," New York, NY, USA, may 2011. [Online]. Available: <https://doi.org/10.1145/1961189.1961199>
- [35] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, Oct 2001. [Online]. Available: <https://doi.org/10.1023/A:1010933404324>
- [36] R. E. Schapire, *Explaining AdaBoost*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 37–52. [Online]. Available: [https://doi.org/10.1007/978-3-642-41136-6\\_5](https://doi.org/10.1007/978-3-642-41136-6_5)

# On-Device Major Indian Language Identification Classifiers to Run on Low Resource Devices

Yashwanth Y S

R.V College of Engineering  
Computer Science and Engineering Department  
Bengaluru, Karnataka, India

**Abstract**—Language Identification acts a first and necessary step in building intelligent Natural Language Processing (NLP) systems that handle code mixed data. There is a lot of work around this problem, but there is still scope for improvement, especially for local Indian languages. Also, earlier works mostly concentrates on just accuracy of the model and neglects the information like, whether they can be used on low resource devices like mobiles and wearable devices like smart watches with considerable latency. Here, this paper discusses about both binary classification and multiclass classification using character grams as the features. Considering total nine languages in this classification which includes, eight code mixed Indian languages with English (Hindi, Bengali, Kannada, Tamil, Telugu, Gujarati, Marathi, Malayalam) and standard English. Binary classifier discussed in this paper will classify Hinglish (Hindi when written using English script is commonly known as Hinglish) from seven other code-mixed Indian Languages with English and standard English. Multiclass classifier will classify the previously mentioned languages. Binary classifier gave an accuracy of 96% on the test data and the size of the model was 1.4 MB and achieved an accuracy of 87% with multiclass classifier on same test set with model size of 3.6 MB.

**Keywords**—Character grams; code-mixed; deep learning; Indian languages; language identification; NLP; social media text

## I. INTRODUCTION

In recent years, there has been a boom in the social media usage, especially in India, because of deep penetration of internet connectivity among people. There are close to half a billion active social media users with a growth of 4.2% every year [1]. Due to this rapid increase, people of various demographics and ages have started to use social media and in turn code mixed language has become more popular than ever on social media. Usually, a local regional language is mixed with English. Be it for hate speech detection on social media or to generate auto reply suggestions to incoming text message or any other NLP system that involves code mixed data, requires language tagging as first step and will determine the accuracy of the system as whole to a great extent. There are many difficulties with language tagging. Even though large amount of code-mixed language data is available to us as raw data (tweets, posts, blogs, etc.) on social media, we don't have a readily available tagged data set suitable for supervised learning [2]. Also, there are many dialects of same language, and so there are different spellings and pronunciations of the same word in code mixed context. Another important difficulty is, if the data set is not diverse, then the supervised language tagging model will suffer from over fitting [3]. This

paper aims in building code mixed Indian languages classifiers that can run on low resource devices with little latency, so that they can be used in real time applications like auto reply systems.

The reason for choosing the previously mentioned nine languages for classification is, 82% of Indian population speak one of the nine languages as their first language [4]. Since Hindi is spoken by 58% of Indian population [4], this article also considers a binary classifier that distinguishes Hinglish from other languages along with the multi class classifier. Collected close to 120k sentences for all the nine languages to create a separate train set and test set for training and testing model respectively.

In this article presents Stacked Bi-LSTM network that uses trigrams as features to classify Hinglish from other languages and Ensemble CNN - Bi-LSTM with attention network with both trigrams and quad grams as features for multiclass classification [5][6][7].

## II. RELATED WORK

In recent years, lot of research on language identification is done, which essentially is the first step in NLP systems, although less work is done where it involves detection of multiple Indian languages. Inumella Chaitanya et al. describe how common word embeddings like Continuous Bag of Words (CBOW) and Skip Grams models can be used to generate embeddings that can be feed to common machine learning models like support vector machine, Logistic Regression and K-Nearest neighbors among other algorithms [8]. Anupam Jamatia et al. in their paper describe about two models i.e., Bi-LSTM classifier and Conditional Random Fields (CRF) classifier and suggest that Bi-LSTM classifier performs better. Ramachandra Joshi and Raviraj Joshi describe about various input representations like character, sub-word and word embeddings, for language identification task in their paper. They also pass these representations as input to CNN and LSTM based models. They indicate that sub-word representation combined with LSTM model gives the best results [9].

Sourya Dipta Das et al. train multiple LSTM models and create an ensemble model using stacking and threshold technique which helping in increasing the accuracy of the model instead of using a single model [10]. Neelakshi Sarma et al. have developed a framework for language identification which is capable of recognizing words that are borrowed from other languages and used in multiple languages and predict



the language. This framework also considers the context of the sentence [11].

Monojit Choudhury et al. used both word embeddings and character embeddings, then concatenated both of them to predict the language [12]. Harsh Jhamtani et al describes a method which combines word embeddings, character n-grams and POS tagging to predict the language [13]. Shruti Rijhwani et al presents a unsupervised model for language detection that does not require any annotated data to train the model. It is also capable of detecting large number of languages [14].

### III. PROPOSED SOLUTION

This section discusses about the dataset creation, pre-processing steps, embeddings generation and finally about the classifiers i.e., binary and multiclass classifiers. In summary, first a data set is created by collecting data from various sources. Vocabulary or Embeddings are generated from the data set, which will be used to encode the input before giving it the model. All the steps are explained in detail, in following sections and summary of the steps in very high level is shown in Fig. 1.

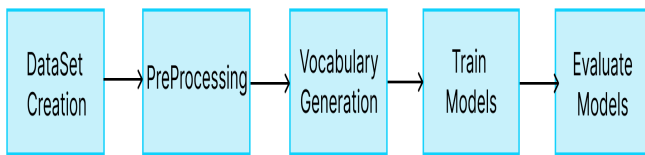


Fig. 1. Summary of Steps of Proposed Solution.

#### A. Dataset

There was no readily available dataset for the previously mentioned nine languages. Had to create one by collecting data from multiple sources like reddit, Facebook posts, twitter tweets, WhatsApp chats and blogs. English language sentences are easy to collect but it is difficult to find and collect other Indian local languages code mixed with English. Data for Indian languages code mixed with English, was collected by scrapping from dedicated subreddit topics and twitter tweets. Also, used Google translate API, where we provide a English sentence as input to translate function of the API and set the destination language to any one of the Indian languages [15]. The translate function returns an object that has a pronunciation attribute. This pronunciation is a close approximation of Indian language sentences written in English script. In addition to this, after collecting the data, some irregularities to the spellings were introduced in the data set, so that the model does not suffer from overfitting problem. In total, the dataset consists of little above 100k sentences and the test set consists of about 13k sentences. The test data consists mostly of real world data and the approximation data generated from google translate API is not included. The distribution of train set and test set is shown in Table I and Table II, respectively.

#### B. Pre-Processing

These common preprocessing steps are applied to all the sentences in dataset and also to the input sentence given to

TABLE I. TRAIN SET

Languages	Number of Sentences
Hinglish	26804
Bengali	11570
Kannada	11148
Tamil	11115
English	11069
Gujarati	11032
Telugu	9231
Marthi	8166
Malayalam	6836
Total	106971

the trained model to predict. Removed the components of the sentences that don't help us in identifying the language. Digits, punctuation marks, extra whitespaces, HTML tags and emojis are removed.

Apart from these common preprocessing steps, irregularities to spellings of Indian languages is introduced into only train set. This is because, Indian language words when written in English script can have different spellings. For instance, the word "why" in Telugu language is spelled as "enduku" or "yenduku", when written in English script. Similarly, "where" in Kannada language is spelled as "ellige" or "yellige". So, introduced these extra letter in some cases into train set so that the model generalizes well and does not suffer from overfitting problem. The summary of preprocessing steps are shown in Fig. 2.

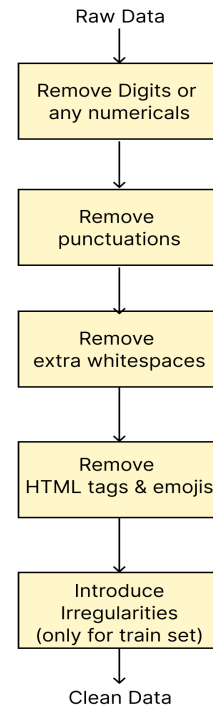


Fig. 2. Summary of Pre-Processing Steps.

TABLE II. TEST SET

Languages	Number of Sentences
Hinglish	2496
Bengali	689
Kannada	1338
Tamil	1554
English	3226
Gujarati	262
Telugu	1599
Marthi	1942
Malayalam	442
Total	13548

### C. Vocabulary/Embeddings Generation

Generated trigrams and quadgrams using all the sentences from the dataset and stored these character grams in form of a dictionary, where the character gram string is the key and a unique integer starting from 1 as the value for the key. So, each character gram is unique identified by an integer. Zero (0) is used to indicate out of vocabulary (OOV) character grams. Only top 20k (in terms of number times it appears in dataset) trigrams and quadgrams are selected. We are limiting the size of the embeddings to limit the size of the model, keeping in mind that we need models that can run on-device with as little resource utilization as possible. Separate dictionaries are created for trigrams and quadgrams. Trigrams are used as embeddings in binary classifier and both trigrams and quadgrams are used in case of multiclass classifier. These dictionaries are stored as csv files.

The process of encoding the input that must be given to the model, using the vocabulary dictionaries generated is described next. For instance, lets consider the following trigram vocabulary dictionary and input sentence in English as shown in Fig. 3. The input sentence is then split into trigrams and each of these trigrams are searched in the vocabulary dictionary. If there is a match, then value associated with the trigram in the dictionary is used for encoding. If the trigram match is not found, then it is a case of out of vocabulary (OOV) and 0 is used to encode such trigrams. Same is done when quadgrams are used as features.

### D. Binary Classifier

After following the preprocessing steps, the sentences in the train set are encoded using the earlier generated embeddings. Each sentence encoded is padded to have a size of 50. The first layer of the model is Embedding layer with vocab size of 20k, as previously mentioned and with both the input and output dimensions to be 50. The activation function for this layer was relu. The embedding layer is followed by a dropout layer (rate = 0.5). This followed by stacked Bi-LSTM layers with output dimensionality of 32 and 16. Stacking the Bi-LSTM layers, helped in boosting the accuracy of the model. This is followed by a dense layer with output dimensionality of 8 followed by drop out layer (rate = 0.5). Last is the output layer with sigmoid as the activation function. The summary of the model is shown in Fig. 4. The entire model was compiled with adam as optimization function and binary cross-entropy

Consider the following example vocabulary and input

```
trigram_vocabulary = [ 'hai':10 , 'kya':20 , 'hel':5 , 'the': 16 ,  
                      'ell':7 , 'llo':10 , 'ere':14 , 'her':8 ]
```

```
input = [ 'hello there' ]
```



Input Sentence is split into trigrams

```
input_trigrams = [ 'hel', 'ell', 'llo', 'lo ', 'o t'  
                  , ' th', 'the', 'her', 'ere' ]
```



See whether vocabulary has the trigram and encode with corresponding value, otherwise 0

```
input_encoding = [ 5 , 7 , 10 , 0 , 0 , 0 , 16 , 8 , 14 ]
```

Fig. 3. Steps to Encode Input.

as loss. The number of epochs was set to 3 and batch size was set to 64.

The trained model is then saved as a TensorFlow Lite (tflite) model in order to compress the size and decrease the latency of the model [16], without losing much on accuracy and tflite models are easier to use on low resource devices like mobiles or even embedded devices. The tflite model size came to be 1.4 MB. The coming sections discuss about the accuracy and performance of the model in detail.

### E. Multiclass Classifier

The preprocessing and encoding steps are same as for binary classifier, except both trigrams and quadgrams are used. Siamese neural networks are used, whose output is concatenated and given to a dense network to form an ensemble model [17][18]. Built two models, with trigrams as features for one and quadgrams as features for the other, then concatenated the outputs and feed it to a deep network as shown in Fig. 5. After the embedding layer we use Conv1D layers with varying kernel sizes (here 3, 4 and 5) with relu as activation function and followed spatial dropout (rate = 0.2) and max pooling cells. Using filters of various sizes, dropout and max pooling cells helped reduce overfitting and variance largely. This is followed by Bi-LSTM stack with attention with 32 and 16 sizes orderly. The output from two identical networks using different features was combined by a concatenate layer followed by two dense layers of 16 and 9 orderly as shown. The entire model was compiled with adam as optimization function and categorical cross-entropy as loss. The number of epochs was set to 3 and batch size was set to 64. Even the multiclass classifier is stored as a tflite model and the size comes to 3.6 MB. The accuracy and performance of the model is discussed in the coming sections.

TABLE III. MODELS PERFORMANCE SUMMARY

Models	Accuracy	F1-score	Model Size	Latency	Peak RAM Usage
Binary Classifier	96%	0.96	1.4MB	8-10ms	4.7MB
Multiclass Classifier	87%	0.88	3.6MB	30-35ms	8.3MB

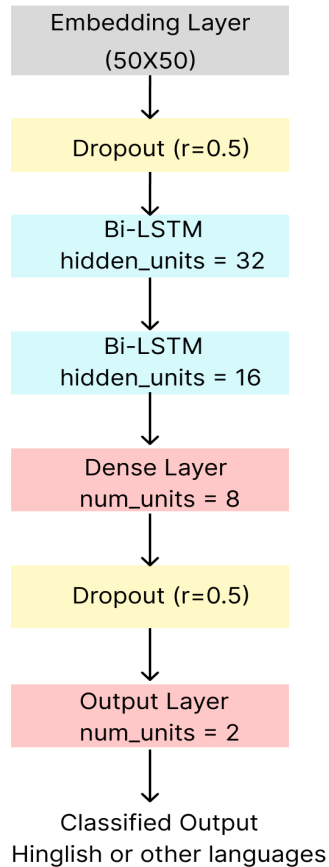


Fig. 4. Binary Classifier Model Summary.

#### IV. EXPERIMENTAL RESULTS

The parameters on which the models was judged are accuracy, f1-score, size, latency and the peak RAM usage by the model. The test set described earlier is used to measure the performance of the model. A c++ program was written to perform the earlier discussed preprocessing steps on the input which is to be given to trained model and to load the saved tflite model and predict the output. The reason for writing c++ code to estimate the performance of the models rather than python was, c++ code runs faster than python [19].

The latency of the model was estimated by recording the time (say t1) when the input is given for preprocessing and again recording the time (say t2) when model predicts the output. The latency is calculated as difference of t2 and t1. So, this latency also includes the time taken for preprocessing steps as well. The now() function of high\_resolution\_clock class as part of chrono c++ header was used to record time t1 and t2 [20]. The range of latency of both the models is recorded in Table III.

TABLE IV. BINARY CLASSIFIER PERFORMANCE FOR EACH CLASS

Languages	Precision	Recall	F1-score
Hinglish	0.95	0.96	0.95
Other Languages	0.98	0.97	0.97

TABLE V. MULTICLASS CLASSIFIER PERFORMANCE FOR EACH CLASS

Languages	Precision	Recall	F1-score
Hinglish	0.92	0.96	0.94
Bengali	0.84	0.98	0.90
Kannada	0.83	0.86	0.85
Tamil	0.87	0.82	0.84
English	0.90	0.90	0.90
Gujarati	0.75	0.62	0.69
Telugu	0.86	0.88	0.87
Marthi	0.98	0.98	0.98
Malayalam	0.67	0.48	0.59

The peak RAM usage was estimated using massif tool, which is heap profiler [21]. Massif comes as part of valgrind, which is collection of tools for memory profiling. The peak usage for both the models is recorded in Table III.

The same vocabulary dictionaries saved as csv files earlier are used to encode the test input sentences given to the trained models to predict the output. The predicted labels given by the trained model is stored and compared with the actual labels, then the classification report is generated based on it. The classification report for both binary classifier and multiclass classifier generated using classification\_report part of scikit-learn package [22] and is reported in Table IV and Table V, respectively.

#### V. ANALYSIS

The binary classifier performs well on most kinds of data. The only drawback observed was, when the input sentence is very short and mostly filled with English, then it misclassifies Hinglish as English. For instance, “Relatives aya” (which means “Relatives came” in English) is classified as English sentence instead Hinglish. This is because the sentence is very short and has only a three letter Hindi word as part of it and the model finds it difficult to predict the correct label. Similarly, “Ek spoon” (which means one spoon) is also misclassified as English.

The multiclass classifier does well on most languages, except Malayalam and Gujarati. The F1-scores for Malayalam and Gujarati classes are 0.59 and 0.69 respectively, which is way less compared to other classes as seen from Table V. These classes also pull down the overall accuracy of the multiclass classifier. The main reason for this is because Malayalam language is very closely related to Tamil Language. Out of 442 Malayalam sentences in the test set, 190 of are missclassified

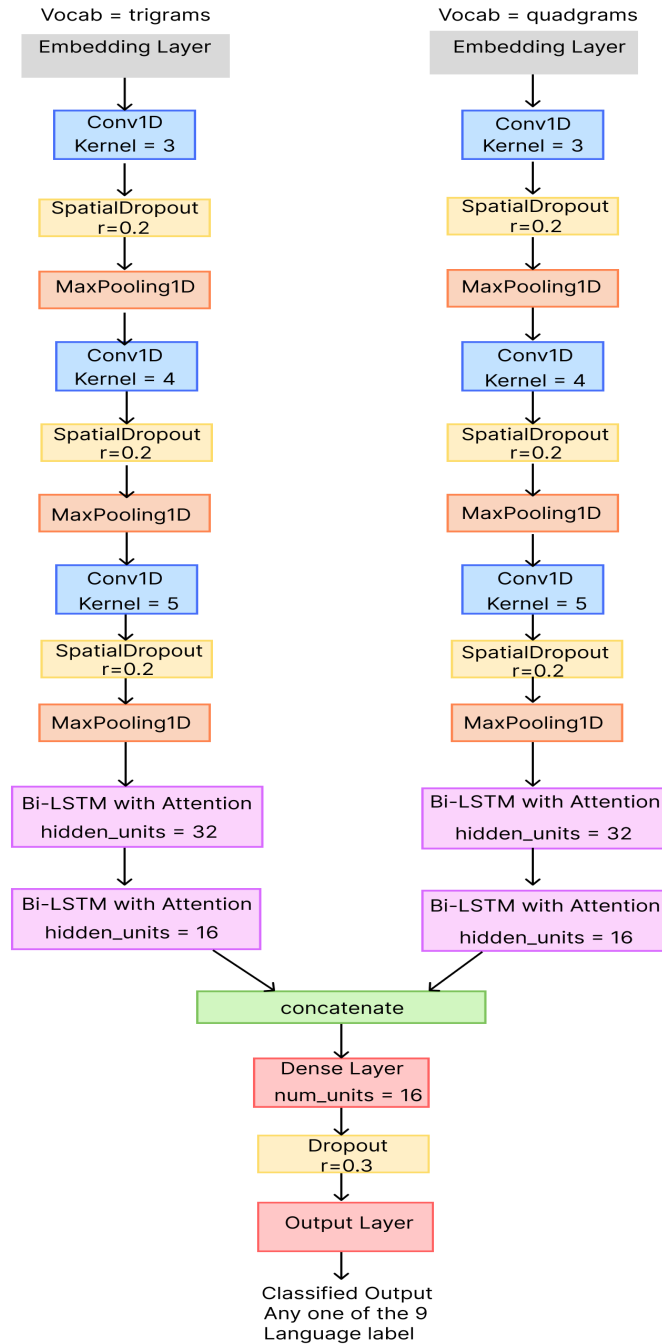


Fig. 5. Multiclass Classifier Model Summary.

as Tamil. These two languages are spoken by people of adjacent states of South India, and the root words of both languages are very similar. For instance, “come home” in Malayalam is “Vittir varu” and in Tamil is “Vittirku va”. The encoded sequences of these two sentences using trigrams or quadgrams will be very close. This is the reason the model confuses for Malayalam sentences and classifies them as Tamil sentences. The same reason goes for Gujarati language class, where Gujarati language is very closely related to Hindi and encoded sequences are also close and the model misclassifies. Out of 262 Gujarati sentences in the test set, 70 are misclassified as

Hinglish.

## VI. CONCLUSION AND FUTURE WORK

The binary classifier performs really on diverse data and generalizes well, expect for really small sentences. The multiclass classifier performs well on seven out of nine languages. With model sizes of 1.4 MB and 3.6 MB for binary classifier and multiclass classifier, these models can be used on any low resource devices like smart watches, mobiles or any embedded system where memory and RAM consumption are a constraint.

The future work would be to collect more real world data to train, in place of approximate data generated using Google translate API. Also, the vocabulary size was restricted to limit the resource usage on the device. So, if the models are run on powerful devices, then the vocabulary dictionaries can be further extended further and models can be retrained with the extended vocabulary and check if there is any effect on accuracy of the models. Also, the embedding layer size and the number of hidden layers of the neural network can be increased to increase the accuracy the model.

## REFERENCES

- [1] <https://www.theglobalstatistics.com/india-social-media-statistics/>
- [2] Osisanwo F.Y., Akinsola J.E.T., Awodele O., Hinmikaiye J. O., Olakanmi O., Akinjobi J. "Supervised Machine Learning Algorithms: Classification and Comparison". International Journal of Computer Trends and Technology (IJCTT) V48(3):128-138, June 2017. ISSN:2231-2803
- [3] <https://machinelearningmastery.com/overfitting-and-underfitting-with-machine-learning-algorithms/>
- [4] [https://en.wikipedia.org/wiki/List\\_of\\_languages\\_by\\_number\\_of\\_native\\_speakers\\_in\\_India](https://en.wikipedia.org/wiki/List_of_languages_by_number_of_native_speakers_in_India)
- [5] Sepp Hochreiter and Jürgen Schmidhuber. "Long Short-Term Memory". Neural Computation. Volume 9, Issue 8. November 15, 1997. pp 1735–1780. <https://doi.org/10.1162/neco.1997.9.8.1735>
- [6] Dong, X., Yu, Z., Cao, W. et al. A survey on ensemble learning. Front. Comput. Sci. 14, 241–258 (2020). <https://doi.org/10.1007/s11704-019-8208-z>
- [7] S. Albawi, T. A. Mohammed and S. Al-Zawi, "Understanding of a convolutional neural network," 2017 International Conference on Engineering and Technology (ICET), 2017, pp. 1-6, doi: 10.1109/ICEngTechnol.2017.8308186.
- [8] Inumella Chaitanya, Indeevar Madapakula, Subham Kumar Gupta and S thara. 2018. "Word Level Language Identification in Code-Mixed Data using Word Embedding Methods for Indian Languages". 2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI). ISBN: 978-1-5386-5214-2
- [9] Joshi, R., Joshi, R. (2022). "Evaluating Input Representation for Language Identification in Hindi-English Code Mixed Text". In: Kumar, A., Senatore, S., Gunjan, V.K. (eds) ICDSMLA 2020. Lecture Notes in Electrical Engineering, vol 783. Springer, Singapore. [https://doi.org/10.1007/978-981-16-3690-5\\_73](https://doi.org/10.1007/978-981-16-3690-5_73)
- [10] Sourya Dipta Das, Soumil Mandal, Dipankar Das. "Language Identification of Bengali-English Code-Mixed Data using Character & Phonetic based LSTM Models". FIRE '19: Proceedings of the 11th Forum for Information Retrieval Evaluation. December 2019. Pages 60–64. <https://doi.org/10.1145/3368567.3368578>
- [11] Neelakshi Sarma, Ranbir Sanasam Singh, Diganta Goswami. "Switch-Net: Learning to switch for word-level language identification in code-mixed social media text". Natural Language Engineering. Volume 28 Issue 3. DOI: 10.1017/s1351324921000115
- [12] Monojit Choudhury, Kalika Bali, Sunayana Sitaram, and Ashutosh Baheti. 2017. "Curriculum design for code-switching: Experiments with language identification and language modeling with deep neural networks". In Proceedings of the 14th International Conference on Natural Language Processing (ICON-2017), pages 65–74, Kolkata, India. NLP Association of India.
- [13] Harsh Jhamtani, Bhogi Suleep Kumar, Vaskar Raychoudhury. "Word-level Language Identification in Bi-lingual Code-switched Texts". Pacific Asia Conference on Language, Information and Computing. December 2014.
- [14] Shruti Rijhwani, Royal Sequiera, Monojit Choudhury, Kalika Bali, and Chandra Shekhar Maddila. 2017. Estimating code-switching on twitter with a novel generalized word-level language detection technique. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), volume 1, pages 1971–1982.
- [15] <https://stackabuse.com/text-translation-with-google-translate-api-in-python/>
- [16] <https://towardsdatascience.com/model-compression-a-look-into-reducing-model-size-8251683c338e>
- [17] <https://towardsdatascience.com/what-are-siamese-neural-networks-in-deep-learning-bb092f749dcb>
- [18] <https://machinelearningmastery.com/tour-of-ensemble-learning-algorithms/>
- [19] <https://towardsdatascience.com/how-fast-is-c-compared-to-python-978f18f474c7>
- [20] [https://en.cppreference.com/w/cpp/chrono/high\\_resolution\\_clock](https://en.cppreference.com/w/cpp/chrono/high_resolution_clock)
- [21] <https://valgrind.org/docs/manual/ms-manual.html>
- [22] <https://scikit-learn.org/stable/>

# Evaluating Hybrid Framework of VASNET and IoT in Disaster Management System

Sia Chiu Shoon<sup>1</sup>, Mohammad Nazim Jambli<sup>2</sup>, Sinarwati Mohamad Suhaili<sup>3</sup>, Nur Haryani Zakaria<sup>4</sup>

Faculty of Computer Science and Information Technology, Universiti Malaysia Sarawak,

94300 Kota Samarahan, Sarawak, Malaysia<sup>1,2</sup>

Pre-University, 94300 Kota Samarahan, Sarawak, Malaysia<sup>3</sup>

School of Computing, College of Arts & Sciences, Universiti Utara Malaysia, 06010 Kedah, Malaysia<sup>4</sup>

**Abstract**—During emergency operations in Disaster Management System (DMS) for natural and man-made disasters, any breakdown in the existing information and communication technology will affect the aspect of effectiveness and efficiency on an emergency response task. For Vehicular Ad-hoc Sensor Network (VASNET), the limitation in terms of infrastructure that consists of RSU (Roadside Sensor Unit) may partially or fully destroy the post-disaster scenario. As such, performance degradation of VASNET affects the network infrastructure on high packet loss, delay, and produce a huge amount of energy consumption in DMS. Thus, modification of VASNET and integrate with Internet of Thing (IoT) technology is a must to improve and solving the current problem on VASNET technology. Therefore, the main objective of this study was to investigate the performance of the proposed modified VASNET framework integrated with IoT at DMS in terms of energy consumption and packet loss. A suggested node in the proposed framework was introduced to implement low data rate and high data rate in evaluating the proposed framework using LTE and LTE-A transmission protocol. It was found that LTE-A contributes more energy by 25.33 (mJ/Byte) compared to LTE on 20 (mJ/Byte) on a high data rate. On the other hand, in terms of low data rate, LTE-A influences the most on the proposed framework by recording 19.82(mJ/Byte), LTE only 19.33 (mJ/Byte). For packet loss, LTE shows a high packet loss rate by contributing 11.39% compared to LTE-A, which is 8.0% in terms of low data rate, and 14.80% compared to LTE-A, only 11.97% for high data rate. Consequently, LTE-A on high data rate contributes more energy consumption and LTE in packet loss on same data rate.

**Keywords**—Energy consumption; packet loss; LTE-A; VASNET; IoT

## I. INTRODUCTION

With the rapid growth of smart devices in Internet of Thing (IoT) technology is estimated that in 2025, the total quantity of smart devices could rise to 1.56 billion globally [1]. As such, research interest on wireless networking communication has focused on reducing energy consumption by applying various methods. Moreover, energy dissipation could pollute our environment caused by electronic devices becoming unhealthy and significant impact on our daily lives.

In connection with this, various natural disasters like landslides, volcanoes and earthquakes may rise from year to year affecting millions of innocent human life [2]. To cope with this problem, research work was done on existing Disaster Management System (DMS) to integrate different types of wireless network technologies to be more efficient. Therefore, the integration of VASNET and IoT in DMS is introduced in

this research. The interface plays the leading role in linking both VASNET and IoT in the DMS [2]. The modification of Vehicular Ad-hoc Sensor Network (VASNET) is hugely challenging in terms of protocol that used to be compatible with current IoT technologies. The modified sensor node from VASNET should be equipped with sensing, processing and transmitting the data [3] concerning Base Station (BS) in Bi-directional mode. Furthermore, the data dissemination must cover a large geographical region [4]. The sensor nodes on disaster areas can capture data in the cluster environment from the tracked region, manipulate data and broadcast to main nodes with more collection points called Gateway, actively using the interface for further data analysis or tracking location [5].

One of the critical resources in sensor nodes that affect the performance and reliability of DMS is the energy supplied [4]. The primary role of power was providing the necessary energy to achieve the mission of sensor nodes typically [6]. A proper energy reduction potentially prolonged the DMS system in terms of stability and lifetime [3]. This enables the Emergency Response Team to save more human life as much as possible [2].

Consequently, energy-saving becomes essential when sensor nodes are powered by their restricted battery. Sensors spread over a large area or in a harsh or hostile area such as volcanoes or even deep-sea when battery power is depleted. It could be difficult or uncomfortable to exchange or recharge the battery [7].

The leading cause of sensor node's energy waste is the radio system [4]. For that reason, several concepts and strategies has been emphasized on power saving in reducing the data sending like scheduling, aggregation, routing and clustering [8]. Generally, when selecting different types of wireless network technologies linking with DMS, a few considerations we need to take into account on the particular application like power consumption and maximum distance range [9].

Overall, this research work greatly benefits the community in terms of minimizing human life as much as possible that contributes to DMS, which is listed below.

- Multichannel
- Network Establishing and Channel Formation
- Interface

Therefore, the principles and structure of this research work are to identify the method proposed in DMS with



modifying VASNET on IoT technologies that determine by using energy consumption and packet loss. This paper was structured as follows: the main challenges on VASNET and IoT are discussed in Section II, related works in Section III, proposed framework in Section IV, performance evaluation and method in Section V, result and discussion in Section VI and conclusion in Section VII.

## II. CHALLENGES ON VASNET AND IOT

There are challenges that cannot be avoided in any wireless network. The primary challenges and limitations in which it could affect the performance in the wireless network are discussed below [3], [10].

- 1) Security
  - Varies method has their strengths and weaknesses and none of them provide the best solution on it. For instance, vehicle-to-vehicle (V2V) focuses on several attributes including authenticity, authority, integrity with confidentiality. Different V2V applications such as e-health systems and smart metering may have various privacy requirements which need to be taken into account during the initial stage of system design [11].
- 2) Mobility
  - High mobility of VASNET for vehicles to move randomly compared to other wireless network infrastructures. Which may contribute a redundant data collected to nearby RSS (Road Side Station) or BS (Base Station).
- 3) Power limitation
  - In VASNET, power constraint is one of the most important challenges in which it shadowed all other aspects like routing, fusion, and the massive battery carried by the device. For instance, car batteries.
- 4) Devices challenges
  - In the same network, it may be equipped with various types of protocol capabilities on different sensors or devices. For instance, a vehicle system is a critical challenge especially in a tracking system as the node is movable [6].
- 5) Big Data
  - A huge amount of data may produce conflicting meanings(vagueness), which requires checking for quality and value. Multiple deployments on similar sensors increase data accuracy. However, it could experience extra noise data [12].
- 6) Other Challenges
  - Several challenges also impact the design of wireless network sensors. For instance, a group of sensor nodes moves into a particular portable robot or various automobiles. It would result in any sensor network topology constantly being altered in which request of changes is repetitive. Some needs for MAC

(Media Access Control) for density modification, routing on neighbour lists modification with data gathering.

## III. RELATED WORKS

The main objective of this review is to thoroughly examine the published works of literature that extend the sensor network regardless of the existing DMS system with VASNET modification in various applications and IoT. It shows a relationship in any sensor work in multiple methods related to this research field. There is limited research involving different types of wireless technologies to be executed in any state-of-art system, which will become very demanding in the future research area.

Below is the discussion of previous works done in any technique, algorithm and method related to the energy consumption in the IoT network field and wireless network.

Energy Harvesting system is introduced in the works of the author [3], in which a clustering algorithm was applied on sensor nodes to form a cluster with Cluster Head (CH). This CH was equipped with an external energy source to supply power to prolong the CH in terms of network efficiency. CH transmit and receive a signal link with BS to enable network communication between the sensor nodes in a particular area. However, this method was restricted when the node movement was dynamic and randomly caused frequent re-clustering and affected the energy supply to be depleted to a certain level.

Another well-known clustering base protocol is Low-Energy Adaptive Clustering Hierarchy (LEACH) in author work [13], has been discussed. In this LEACH, the sensor nodes were organised into the cluster, with each cluster are randomly selected. The weak point of this LEACH is that CH was experienced less residual energy on selected CH, which would result in inactive mode quickly. As such, the whole cluster would fall into a non-functional mode and reduce the effectiveness of this LEACH clustering protocol.

To reduce energy consumption on wireless networking for IoT, the author [4] proposed a data reduction method that works on Gateway of network level. It operates on a group of data received by identifying and removing the redundant data set that undergoes a classification process. The author suggested a clustering candidate set to verify a similarity among the members. The data sets after clustering candidate sets were able to transmit through Gateway with minimum energy consumption. To some extent, this research work may have risks when the same valuable data sets have been removed and time-consuming while facing many sensor nodes.

The compression algorithm applied from one data form of sensor node was installed in a wireless sensor network in an underground tunnel in terms of spatial-temporal data has been explored by the author [14]. The proposed algorithm effectively operates on temporal and spatial characteristics for the sensor's data. The data recovery method was nearly approximate to an initial data node. This algorithm served a high complexity for data compression and, therefore, was difficult to be considered on the limited resources node practice in IoT networks.

The idea of the Prefix-Frequency Filtering (PFF) technique was proposed by the author [15]. This technique is divided

into two phases; the sensor phase to utilize the local data processing and the aggregate stage for PFF with Jaccard Similarity mechanism that can consolidate data similar from nearby sensor nodes. However, this technique was not competitive in reducing redundant data before broadcasting to the BS.

Implementation of decentralized hierarchical clustering is proposed to avoid the redundant control message transmitted from sensor nodes to BS done by the author in [10]. The formation of sensor nodes clusters with a criterion of intra-cluster among sensor nodes and CH being selected according to the shortest distance between sensor nodes and BS. CH was revalidated each time on sensor nodes remaining energy. Thus, it identified energy depletion and overloading on any particular sensor nodes. But when operating in multiple criteria, this algorithm works in performance-less that includes heavy computational complexity in the data transferring process.

The author introduced an Adaptive Lossless Data Compression (ALDC) algorithm in [16] for wireless sensor networks. It incorporates several coding to achieve lossless compression alternatively. This method permits the adjustment of compression dynamically to the changeable source. It comprises blocks, and each block implements the optimal compression method. However, this method increases the time complexity and is not suitable to apply on the gateway.

#### IV. PROPOSED FRAMEWORK

In order to justify the reason to conduct this research work. The current VASNET and IoT framework underwent a study to identify the problem and limitations. RSU (Road Side Unit) linking with a gateway for VASNET networking is a disadvantage during disaster or catastrophe occur. RSU may partially or fully destroy during the impact of the disaster, especially by tsunamis and landslides. Communication breakdown in the affected area can cause survivors trapped inside the vehicle to face difficulty to be rescued by ERT members. High node density of vehicles in one area and shorten communication range can be a reason or problem for the existing VASNET framework. Packet congestion or high traffic on VASNET degrade the performance of data transmission and retrieve data that might be experienced by ERT members. On the other hand for the IoT framework. It will be costly to equip a device which able to sustain an extreme situation in high temperature and high-pressure environment, regardless of communication range or protocol. High maintenance of IoT networking is one of the main reasons, why IoT is not suitable to be implemented on existing DMS. It involves a large area, for instance, the metropolitan city which causes a million of financial support that not all the country can handle like the third-world country. As a result of this, modification of VASNET is a must to overcome the short-distance communication coverage and reduce network congestion by using IoT technology. IoT can suit this VASNET to implement point-to-point network connection to cost down current IoT infrastructure. As such, this research aims to combine and modify both VASNET and IoT to become more stable and reliable to apply to DMS systems. Consequently, the interface is introduced in this research study, which is a medium to link different protocols on VASNET and IoT through varying artificial intelligent devices like smartphones and smart switches through the algorithm proposed.

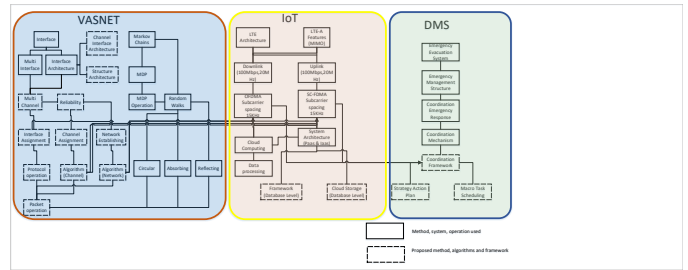


Fig. 1. Proposed Hybrid Framework.

The most complex and challenging part is integrating two different network topologies into a DMS system that enhances performance in terms of stability and reliability. The interface becomes the key that enables VASNET and IoT to communicate simultaneously. Thus, execution of both wireless networking potentially generates an unwanted energy dissipation which causes degradation of hybrid network performance. To solve this problem, the transmission medium and modulation module becomes crucial as the purpose is to reduce energy consumption as minimum as possible. Therefore, the proposed framework depicted in Fig. 1 illustrated this hybrid model implemented on the DMS system.

It can be observed from the diagram in Fig. 1, existing VASNET has to be modified by introducing an interface with Channel Assignment and Network Establishing algorithms. Those algorithms have been implemented by applying Markov Chains and LTE with LTE-A to configure a nearby node in a post-disaster scenario. Markov Chains potentially locate the latest state of survivor that depends on the existing state especially in heavy floods and volcano disasters, making the movement of humans trapped in random mode in any emergency. Therefore, Markov Chains implementation can connect and link all nodes (Survivor) in a certain disaster area. Furthermore, LTE and LTE-A are suggested as a communication protocols with less latency and coverage area up to 5km. With this, it enables the Emergency Respond Team (ERT) to connect with survivors without any barriers in terms of communication. This algorithm plays an important role to integrate with IoT structure and the idea of channel interface and structure architecture on interface able to link both VASNET and IoT together. The objective of this interface is designed for a multichannel or multimode approach. It permits the system to detect and identify as many victim locations (node) in the short and medium range.

This interface connects the VASNET and IoT architecture on particular Long Term Evolution (LTE) for Downlink and Uplink medium to optimize the communication flow. The LTE architecture in IoT can handle multi-node that result in shortened latencies in terms of signal control [17]. Moreover, IoT network architecture is upgraded to add in suggested Database in Cloud environment to the IoT architecture due with Platform as a Service (PaaS) and Infrastructure as a Service (IaaS) level. The advantage of IoT in the Multi-node approach is that it works on OFDMA in Downlink and SC-FDMA on Uplink, which is the main critical point on Downlink. It could contribute more energy when ERT (Emergency Response Team) communicates with various victims to locate the victim more accurately. OFDMA has the characteristics

of separating the data into several narrowband subcarriers to improve the bandwidth [17] during emergency periods.

The hybrid architecture would be connected to existing DMS to enhance the DMS performance in which the DMS system should equip with the coordination framework. This coordination framework is explored to suit the hybrid framework that can receive valuable data or information in a real-time situation during an emergency. Strategy Action Plan and Macro Task Scheduling is the idea to optimum the DMS system performance link with this hybrid framework, and capable to strategy the rescue plan effectively to rescue innocent people during disaster or catastrophe occurs.

### V. PERFORMANCE EVALUATION METHOD

To examine the performance of a hybrid framework in terms of stability and reliability. The energy consumption on BS in this hybrid framework will be determined as this is the most essential and significant part. As a result, the parameter of energy consumption and packet loss at BS will undergo testing, and all the evaluation methods or steps are elaborated on below.

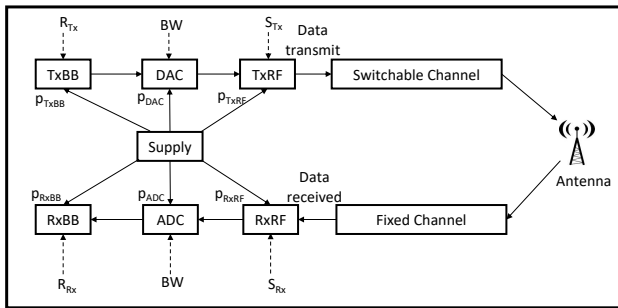


Fig. 2. The Energy Consumption Model.

Fig. 2 depicts the power consumption model node consisting of power level (s), transmit  $T_x$  and receiver  $R_x$ , Base Band (BB), Bandwidth (BW) and radio frequency  $T_x$  and  $R_x$ . Where P is classified as total power consumption in idle and connected mode,  $R_{xRF}$  and  $T_{xRF}$  in RF part on  $R_x$  and  $T_x$  chains are consumption on each other.  $R_{xBB}$  and  $T_{xBB}$  are consumed on BB parts, and 2CW is increased while two codewords (CW) are in the downlink. For parameters on  $P_{Rx}$ ,  $P_{Tx}$  and  $P_{Rx+Tx}$  are on idle, receive, transmit, and use 2CW. The  $R_x$  and  $T_x$  power level is considered S and  $R_x$ , and  $T_x$  is the R data rate individually.

Each respective node was set to 50mW with low power consumption to obey LTE-A characteristics. Each packet converted was divided into power settings concerning the base station to gather each node's energy consumption.

$$\sum_{i=1}^n P_T = m_{idle(i \rightarrow n)} P_{idle(i \rightarrow n)} + \overline{m_{idle(i \rightarrow n)}} \{ P_{con} + m_{Tx} \cdot m_{Rx} \cdot P_{Tx+Rx} + m_{Rx} [P_{Rx} + P_{RxRF}(SR_x) + P_{RxBB}(R_{Rx}) + m_{2cw} \cdot P_{2cw}] \cdot m_{Tx} [P_{Tx} + P_{TxRF} + P_{TxBB} R_{Tx}] \} w \quad (1)$$

$$\sum_{i=1}^n P_T = \frac{\sum_{i=1}^n \text{Joule}}{\text{Seconds}} \quad (2)$$

$$\sum_{i=1}^n \text{Joule} = \sum_{i=1}^n P_T \times \text{Second}(s) \quad (3)$$

$$\begin{aligned} \sum_{i=1}^n \text{EnergyConsumption(Average)} &= \frac{\sum_{i=1}^n \text{Joule}}{\text{Byte}} \\ &= \frac{\sum_{i=1}^n (P_t \times \text{Second}(s))}{\text{Byte}} \quad (4) \end{aligned}$$

We apply LTE on the same hybrid framework to compare the LTE-A energy consumption.

### VI. RESULT AND DISCUSSION

We consider downlink as principles measurement for energy consumption as it is the most significant to contribute energy consumed on network access that can evaluate energy efficiency [18]. Furthermore, low and high data rates have also been used to determine the proposed framework in parameter CQI (Channel Quality Indication), which is CQI=2 (low) and CQI =7 (high). Below is the tabular form and graphical form of this experiment result obtained.

TABLE I. ENERGY CONSUMPTION FOR LOW DATA RATE ON 500M FOR 15s

Node	LTE	LTE-A
20	25.00	10.00
40	64.00	60.00
60	75.00	71.43
80	80.00	77.78
100	83.33	79.82

TABLE II. ENERGY CONSUMPTION FOR HIGH DATA RATE ON 500M FOR 15s

Node	LTE	LTE-A
20	50.00	33.33
40	75.00	66.67
60	85.00	80.00
80	90.00	87.00
100	95.00	92.00

Table I and II show the result for low data rate (CQI=2) and high data rate (CQI=7) on 500m and 15s.

Fig. 3 shows energy consumption for low data rate (CQI=2) and high data rate (CQI=7) for LTE and LTE-A with 500m range.

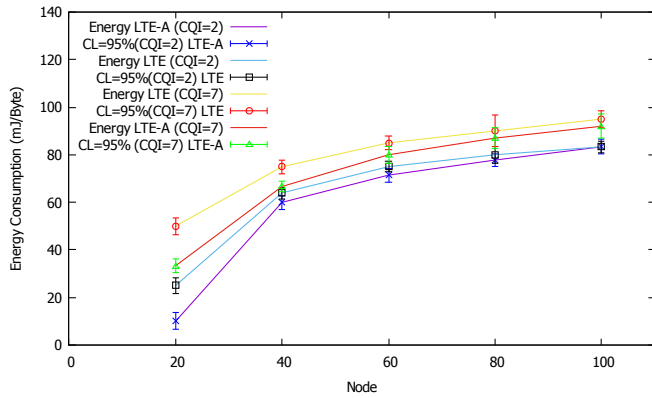


Fig. 3. Energy Consumption Result for 500m on 15s for High Data Rate (CQI=7) and Low Data Rate (CQI=2).

As presented in Fig. 3, the most effective increasing rate was at node 40 group that was showing high data rate for LTE contributes 25 (mJ/Byte), LTE-A 33.34 (mJ/Byte). LTE contributes 39 (mJ/Byte) for low data rates and LTE-A 50 (mJ/Byte). On node 40 onward, all LTE and LTE-A gradually increase, reaching LTE as 20 (mJ/Byte) for a high data rate, LTE-A 25.33 (mJ/Byte), respectively. LTE was 19.33 (mJ/Byte) and LTE-A for 19.82 (mJ/Byte) on low data rate. This is because network formation and channel establishment executed at an initial stage establish a connection among central stations with all the nodes nearby. Besides that, the high mobility of nodes could contribute to the amount of energy consumption as packet loss happens concurrently. Retransmission of packet needs more energy, and it keeps increasing with the condition of the number of nodes in increasing trend and various pattern change also increase simultaneously. Consequently, the channel quality reflected the amount of energy dissipation and consumed more on high data rate compared to the low data rate.

TABLE III. PACKET LOSS FOR LOW DATA RATE ON 300M

Node	LTE	LTE-A
20	8.87E-03	8.74E-03
40	9.58E-03	9.50E-03
60	9.88E-03	9.80E-03
80	9.89E-03	9.85E-03
100	9.96E-03	9.90E-03

TABLE IV. PACKET LOSS FOR HIGH DATA RATE ON 300M

Node	LTE	LTE-A
20	8.87E-03	8.87E-03
40	9.42E-03	9.42E-03
60	9.90E-03	9.73E-03
80	1.00E-02	9.83E-03
100	1.02E-02	9.94E-03

Table III and IV show the experiment results for packet loss for LTE and LTE-A for high data rate (CQI=7) and low data rate (CQI=2).

Fig. 4 and 5 show the tremendous packet loss increase fell into node 40 group for low data rate on LTE 11.39% and

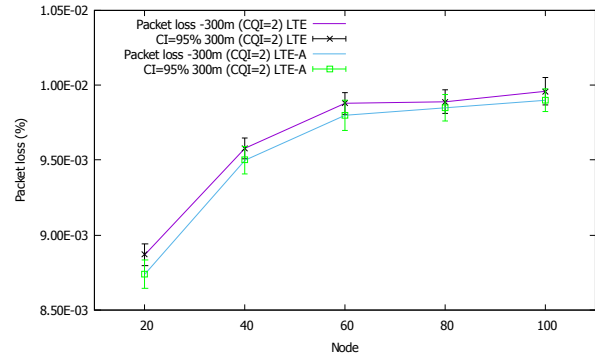


Fig. 4. LTE-A and LTE Packet Loss for 300m (CQI=2).

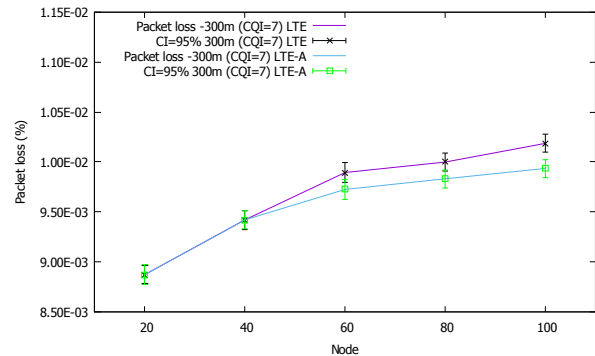


Fig. 5. LTE-A and LTE Packet Loss for 300m (CQI=7).

LTE-A 8.0%. On the other hand, the high data rate presented 14.80% for LTE and 11.97% for LTE-A. Generally, LTE-A improved compared to LTE by about 19.12% for the 300m range.

The packet loss rate gradually increased for both LTE and LTE-A due to the random movement for the 300m range. High-speed mobility can cause the packet transmitted to lose its path because the transition node or relay is out of the network coverage as the central station has to reschedule and retransmit to any particular destination node. Therefore, the existing network has to reform with the neighbour's node to create a new channel and coverage. Instead, a high data rate (CQI=7) experienced a massive amount of energy needed to broadcast the packet in the node-by-node primary. It will produce a lot of energy consumption directly to the network coverage.

## VII. CONCLUSION

In this research work, the modification of VASNET integrated with IoT on DMS was successfully evaluated using low and high data rates. It was able to investigate the effect of stability and reliability by parameters applied to the proposed framework by using energy consumption with packet loss rate.

The result presented a high data rate showing significant influence with more energy on LTE-A as 25.33 (mJ/Byte) compared to LTE 20 (mJ/Byte). At low data rate, LTE-A also gives ultimate contribution by 19.82 (mJ/Byte) higher than LTE with 19.33 (mJ/Byte). The works of packet loss

determined that LTE contributed 11.39% higher than LTE-A on 8.0% for low data rate, and high data rate, LTE is also higher than LTE-A with 14.80% and 11.97%, respectively. Overall, it can be concluded that a high with low data rate on energy consumption parameter, LTE-A significantly impacts proposed framework performance. However, LTE on high and low data rate was higher in terms of packet loss parameter.

Therefore, for further work, 5G is suggested to evaluate the proposed framework in energy consumption measurement and packet loss.

#### ACKNOWLEDGMENT

This work is supported by Universiti Malaysia Sarawak (UNIMAS). The authors would also like to thank the UNIMAS for providing the funds and resources used in this research work.

#### REFERENCES

- [1] F. Al-Turjman, "5g-enabled devices and smart-spaces in social-iot: an overview," *Future Generation Computer Systems*, vol. 92, pp. 732–744, 2019.
- [2] M. N. Jambli, A. S. Khan, and S. C. Shoon, "A survey of vasnet framework to provide infrastructure-less green iots communications for data dissemination in search and rescue operations," *Journal of Electronic Science and Technology*, vol. 14, no. 3, pp. 220–228, 2016.
- [3] A. Rashid, F. Khan, T. Gul, S. Khan, and F. Khan, "Improving energy conservation in wireless sensor networks using energy harvesting system," *Int. J. Adv. Comput. Sci. Appl.*, vol. 9, no. 1, 2018.
- [4] S. A. Abdulzahra, "Energy conservation approach of wireless sensor networks for iot applications," *Karbala International Journal of Modern Science*, vol. 7, no. 4, 2021.
- [5] A. K. Idrees and A. K. M. Al-Qurabat, "Energy-efficient data transmission and aggregation protocol in periodic sensor networks based fog computing," *Journal of Network and Systems Management*, vol. 29, no. 1, pp. 1–24, 2021.
- [6] M. Younan, E. H. Houssein, M. Elhoseny, and A. E.-m. Ali, "Performance analysis for similarity data fusion model for enabling time series indexing in internet of things applications," *PeerJ Computer Science*, vol. 7, p. e500, 2021.
- [7] N. B. Jarah, "Technique pair of node to provide power in wsns," *Karbala International Journal of Modern Science*, vol. 6, no. 2, p. 2, 2020.
- [8] K. Das, S. Das, and A. Mohapatra, "A novel energy-efficient sensor cloud model using data prediction and forecasting techniques," *Karbala International Journal of Modern Science*, vol. 6, no. 3, pp. 1–10, 2020.
- [9] M. S. Mahmoud and A. A. Mohamad, "A study of efficient power consumption wireless communication techniques/modules for internet of things (iot) applications," 2016.
- [10] G. Asha *et al.*, "Energy efficient clustering and routing in a wireless sensor networks," *Procedia computer science*, vol. 134, pp. 178–185, 2018.
- [11] F. K. Banaseka and S. Dotse, "New developments and research challenges for 5g wireless systems and networks," *International Journal of Current Research*, vol. 9, no. 2, pp. 46 626–46 631, 2017.
- [12] M. Mohammadi, A. Al-Fuqaha, S. Sorour, and M. Guizani, "Deep learning for iot big data and streaming analytics: A survey," *IEEE Communications Surveys & Tutorials*, vol. 20, no. 4, pp. 2923–2960, 2018.
- [13] M. Usman, Z. Xing, H. Chiroma, A. Gital, A. Abubakar, A. Usman, and T. Herawan, "Modified low energy adaptive clustering hierarchy protocol for efficient energy consumption in wireless sensor networks," *International Review on Computers and Software*, vol. 9, no. 11, pp. 1904–1915, 2018.
- [14] B. He, Y. Li, H. Huang, and H. Tang, "Spatial-temporal compression and recovery in a wireless sensor network in an underground tunnel environment," *Knowledge and information systems*, vol. 41, no. 2, pp. 449–465, 2014.
- [15] J. M. Bahi, A. Makhoul, and M. Medlej, "A two tiers data aggregation scheme for periodic sensor networks," *Adhoc & Sensor Wireless Networks*, vol. 21, no. 1, 2014.
- [16] J. G. Kolo, S. A. Shanmugam, D. W. G. Lim, L.-M. Ang, and K. P. Seng, "An adaptive lossless data compression scheme for wireless sensor networks," *Journal of Sensors*, vol. 2012, 2012.
- [17] Korhonen, *Introduction to 4G Mobile Communications*, M. Norwood, Ed. Artech House, Inc., 2014.
- [18] M. F. Alotaibi and D. M. Ibrahim, "Coordination emergency response framework (cerf): a proposed model," *International Journal of Computer Science and Information Security (IJCSIS)*, vol. 16, no. 11, 2018.

# A Novel Machine Learning-based Framework for Detecting Religious Arabic Hatred Speech in Social Networks

Mahmoud Masadeh<sup>1</sup>, Hanumanthappa Jayappa Davanager<sup>2</sup> and Abdullah Y. Muaad<sup>3</sup>

Computer Engineering Department, Yarmouk University, Irbid 21163, Jordan<sup>1</sup>

Department of Studies in Computer Science, Mysore University, Manasagangothri  
Mysore 570006, India<sup>2,3</sup>

Sana'a Community College, Sana'a 5695, Yemen<sup>3</sup>

Corresponding Author<sup>3</sup>

**Abstract**—Social media platforms generate a huge amount of data every day. However, liberty of speech through these networks could easily help in spreading hatred. Hate speech is a severe concern endangering the cohesion and structure of civil societies. With the increase in hate and sarcasm among the people who contact others over the internet in this era, there is a dire need for utilizing artificial intelligence (AI) technology innovation that would face this problem. The rampant spread of hate can dangerously break society and severely damage marginalized people or groups. Thus, the identification of hate speech is essential and becoming more challenging, where the recognition of hate speech on time is crucial in stopping its dissemination. The capacity of the Arabic morphology and the scarcity of resources for the Arabic language makes the task of distinguishing hate speech even more demanding. For fast identification of Arabic hate speech in social network comments, this work presents a comprehensive framework with eight machine learning (ML) and deep learning (DL) algorithms, namely Gradient Boosting (GB), K-Nearest Neighbor (K-NN), Logistic Regression (LR), Naive Bayes (NB), Passive Aggressive Classifier (PAC), Support Vector Machine (SVM), Ara-BERT, and BERT-AJGT are implemented. Two representation techniques have been used in the proposed framework in order to extract features: a bag of words followed by BERT-based context text representations. Based on the result and discussion part, context text representation techniques with Ara-BERT and BERT-AJGT outperform all other ML models and related work with accuracy equal to 79% for both models.

**Keywords**—Machine learning; Arabic language; hatred detection; social network; classification algorithm

## I. INTRODUCTION

Low-resource languages, e.g., Arabic, Hindi, and Urdu, do not have a considerable amount of data for training and building conversational Artificial Intelligence (AI) systems. The Arabic language is the authorized language for 22 Arabic countries with roughly more than 422 million aboriginal speakers [1]. Additionally, it is a religious language spoken by more than 1.5 billion Muslims. There are three main types of it: *i*) Classical Arabic which is the language of the Holy Quran, *ii*) Modern Standard Arabic (MSA) which is used by academia, and *iii*) Dialectal Arabic which differs between regions since it is used for daily life networking [2]. Thus, the Arabic language, with a large number of speakers worldwide, is challenging task when we work with AI systems. Moreover, the Arabic language is identified as the 4<sup>th</sup> in the usage on

the internet [3]. However, the sophistication of the Arabic language makes the automatic identification of Arabic hate speech a complex task. Dialectal Arabic doesn't have formal grammar or spelling regulations. Moreover, spelled words can have different importance based on various dialects, which augments the vagueness of the language [4].

Sociable applications, e.g., Facebook, YouTube, and Twitter, are generating an extensive quantity of data which is considered a valuable goldmine for researchers. Social media-generated data helps in recognizing unlawful behavior, restricting potential hurt, and maintaining residents safe [2]. Some users utilize the wild adoption of online social networks to spread radical and biased statements that diffuse hate speech. *Sentiment analysis* (SA), i.e., opinion mining, analyses individuals' thoughts, attitudes, emotions, and opinions towards an entity, e.g., person, object, or service. The SA is implemented at various levels of granularity [5]: *i*) *document-level*: each text fragment is considered as a component with an opinion towards a single object. It aims to categorize a review as positive, negative, or neutral, *ii*) *sentence-level*: aims to extract opinions for a smaller text which is more challenging than SA for a document, and *iii*) *aspect-level*: determines the major features of a belief while focusing on aspect extracting and feeling categorization of aspects. The approaches of semantic analysis could be supervised, unsupervised, or hybrid. For *unsupervised* methods, numerous sentiment phrases are required to reveal the semantic orientation of texts. Thus, lexicon-based approaches are used. However, supervised techniques rely mostly on utilizing data mining tools to build a learning algorithm on a collection of tagged information. A prime application of SA is hate speech detection through online social network [6].

Hate speech is any class of inappropriate language, e.g., insults, slurs, threats, encouraging violence, and impolite language, that targets individuals or groups based on typical attributes such as nationality, religion, ideology, disability, social class, or gender. Hate speech includes racism, misogyny, religious discrimination, and abusive speech. *Racism* implies hate speech that attacks people based on their skin color, race, origin, class, or nationality [7]. *Misogyny* is the hate speech that targets females, i.e., women or girls [8]. Religious bias is hatred vocabulary towards somebody based on their beliefs, faiths, practices, or even the deficiency of religious faiths.



The *abusive speech* represents disrespectful, rude, or criticizing speech to hurt or deliver harmful sentiments.

Hate speech (HS) detection is a branch of offensive language detection. There is an increasing studies for abusive/HS detection for English language. However, it is still very limited for Arabic dialects due to the scarcity of the publicly obtainable resources required for abusive/HS detection in Arabic social media texts. The authors of [9] declared that the harmful online content on social media can be grouped into various categories including: Vicious, Vulgar, Offensive, Violent, Adult content, Terrorism and Spiritual hate speech.

This work targets *religious hate speech* (RHS) that could be insulting, abusive, or hateful. RHS aims to instigate hate, intolerance, or roughness toward people because of their religious faiths. The recent immense usage of social networks mandates applying different *text processing* on such cyberspace. The remarkable amount of generated data requires applying new monitoring tasks such as cyberbullying recognition [10], hate speech detection [6], irony identification [11], and discovery of offensive language [12]. Accordingly, battling hate speech mandates generating and elucidating a considerable amount of data for automatic hatred speech identification by building artificial intelligence-based models, i.e., ML and DL [13].

Lately, detecting abusive and hateful speech has gained increasing attraction from investigators in NLP and computational social sciences societies. Thus, detecting abusive speech and hate speech is essential for online safety. lately, various studies indicated that the existence of hate speech may be related to hate crimes [8]. Therefore, this work aims to enhance the detection of offensive language and hate speech on Arabic text. Detecting religious hate speech in any language, including Arabic, has different challenges, including : 1) the gigantic volume of the data generated over social networks makes it difficult to locate typical patterns and trends in the data, 2) noise may exist in the data, e.g., inaccurate grammar, misspelled phrases, Internet slang, abbreviations, lengthening of words, and multi-lingual scripts, 3) the comments being written in poorly text, and including paralinguistic signs, e.g., emoticons, and hashtags. Moreover, hate detection is a context-dependent task, and it is still missing a consense of what is forming hate speech due to the different cultures, customs and traditions, and 4) since the social networks prevent posting illegal content, users post information that looks authentic and simple but quietly causes a hate speech. Thus, building a tool for the automatic detection of hate speech would be complex [6].

Social platforms, e.g., Twitter and Facebook began battling online hate speech by explaining procedures that limit the use of violent and dehumanizing languages [14]. Moreover, various Arabic countries, where their users of social media sites are adding Arabic content, modified their laws to combat cybercrimes including hate speech. For example, Jordan added a new cybercrime laws [15] that defines hate speech as any action, writing, or speech planned to cause and raise ethical conflict or call for violence and provocation to fighting between the diverse segments of the nation. Regarding the Arabic language, there is a clear shortage in the conducted research for hate speech on online social networks. Thus, artificial intelligence, data mining, and machine learning techniques could be utilized to efficiently perform more research and

experiments on hate speech detection which constitutes a fertile resource for investigation. This work aims to design a prototype for the automatic identification of abusive and hate speech using various ML and DL techniques with a standard data set.

The remaining sections are organized as follows. Section II presents preliminaries necessary to understand the context of the work. Section III highlights some of the important related work. The various aspects of the proposed methodology are explained in Section IV. Section V introduces the experimental setup and analysis. The obtained results and their explanation are discussed in Section VI. Finally, Section VII concludes the paper with future directions.

## II. PRELIMINARIES

### A. Natural Language Processing (NLP)

Natural Language Processing is a major component of Artificial Intelligence (AI). It enables robots to analyse and comprehend human language, enabling them to carry out repetitive activities without human intervention. Machines can analyse and comprehend human language through a process known as NLP. NLP-based approaches process a considerable amount of data to obtain useful knowledge. For that different data mining and machine learning approaches are used. Thus, text pre-processing should be applied in order to prepare text for further processing such as representation features engineering that are required to extract features and pass it to ML approaches. For example, pre-processing could include text tokenization, and stop-word removal.

### B. Machine Learning (ML) Algorithms

ML is used in various applications, e.g., healthcare [16], hardware design [17], quality control [18], and NLP, where this work targets NLP application. Information is an organised collection of discrete pieces of data, and it conceals the whole spectrum of representational patterns. The machine's primary objective is to extract patterns that reflect a certain event. If the machine is able to recognise these patterns, then machine learning has taken place. It demonstrate that by adding fresh data or information, where the computer can make accurate predictions. The authors of [19] have mentioned that the advancements in machine learning especially deep learning enable us to design algorithms that use real-world information to make decisions that seem subjective. As shown in Section IV-B, there are different methods to prepare text for further processing. *Text tokenization*, which is also called text segmentation or lexical analysis, groups the text into tokens/words separated by space. Stop-words such as articles (e.g., a, an, the), conjunctions (e.g., and, but, if), and prepositions (e.g., in, at, on) [20], do not represent a specific meaning. Thus, they should be eliminated. Features in ML are essentially numerical attributes. However, the data may not contain numerical attributes, such as in sentiment analysis. Thus, various types of features (e.g., word, character, so on) are converted into numerical features where such operation is called representation and choosing from them which make ML working properly is called *feature engineering* (*feature selection and feature extraction*).

### C. Hate Speech

Recently, the broad usability of smartphones and the high availability of internet access increased the number of users on social media. Moreover, the rapid growth of social media has made it practically unattainable to manually monitor and inspect the massive amount of messages published online every day. Also, social media witnessed a substantial increase in hate and abusive speech, which is a severe problem worldwide that threatens the solidarity of civil communities. Therefore, automatic detection for hate speech, utilizing various classification techniques, is required to filter such harmful content. Twitter is one of the most importing social media platform which is ubiquitous, informal, and unstructured at the same time. Tweets usually have abbreviations, acronyms, spelling errors, and non-ideal punctuation so designing a model to handle this will be an interesting topic for future work.

### D. Transfer Learning (TL)

ML still has some constraints for specific real-world domains. For example, the requirement of having a tremendous amount of training data which have a distribution similar to the testing data could be difficult to satisfy [21]. Thus, semi-supervised learning could be utilized due to the shortage of labeled data. However, for a small amount of unlabeled data, the build model would be defective. Therefore, transfer learning is a promising procedure for such systems. Transfer learning (TL) is a branch of machine learning (ML) which aims to improve the performance of target learners on specific fields by transferring the knowledge possessed in separate but connected source domains [21]. Thus, constructing target learners will have a reduced dependency on a large number of target-domain data. In ML models, knowledge is not retained or accumulated, where learning is performed without considering past learned knowledge in other tasks. However, in transfer learning, the learning process can be faster, more accurate, and require less training data. TL can be classified into: 1) homogeneous where the disciplines are of the identical feature space, 2) heterogeneous where the disciplines have diverse feature spaces.

### E. Data Oversampling and Undersampling (Re-Sampling)

With the tremendous increase in the size of the generated data in various applications, there is a lack of equality in the labeled data. However, various ML techniques assume equal distribution for the target classes which is not always a realistic assumption. Such class imbalance problems will have a good accuracy while other evaluation metrics including precision, recall, F1-score, and ROC (Receiver Operating Characteristics) score, will not have enough scores. As shown in Fig. 1, Re-sampling including under-sampling or oversampling could be used to resolve the problem of an imbalanced data set. Under-sampling reduces the amount of the majority target samples. On the other hand, oversampling raises the quantity of minority class instances by yielding new instances or reproducing some instances [22].

## III. RELATED WORK

Various researches have been conducted to detect hate speech as a wide notion with different types in the English language. Many proposed works performed hate speech



Fig. 1. Undersampling vs Oversampling [22].

detection as a binary classification problem and considered a broad concept such as detecting bullying and derogatory language. In [23], the authors presented an original technique to detect hatred speech in English tweets. For that, they utilized three models, i.e., logistic regression (LR), XGBoost classifier (XGB), and support vector machine (SVM). The obtained performance showed competitive results compared to standard stacking, base classifiers, and majority voting techniques. The authors of [24] determined and discussed challenges encountered by online automatic techniques for hate speech detection in text. The limited availability of the data, sensitivity in language, and the exact definition of what forms of hate speech are well-known challenges. They proposed a SVM technique with high performance while the decisions are easier to interpret than neural methods. However, the used datasets did not include Arabic text.

In [25], the authors used different machine learning algorithms for the automatic identification of hate speech in tweets written in the Indonesian language. Their results showed that the Multinomial Naive Bayes algorithm has the most promising results with a value of 71.2% and 93.2% for accuracy and recall, respectively. The authors of [2] researched the capability of deep learning based on Convolutional Neural Networks (CNN), CNN-long short-term memory networks (CNN-LSTM), and bidirectional LSTM (BiLSTM-CNN) to automatically detect hateful content posted on social media. For that, they used the ArHS dataset with 9833 tweets, which is believed to be the largest Arabic dataset with hate speech content.

The authors of [14] aimed to identify Cyber hate speech within the Arabic content of Twitter where they used various NLP and ML techniques. In [26], the authors used Twitter to construct an Arabic text detection hate speech model. They use this knowledge to analyze a dataset of 11 thousand tweets. They apply the Term Frequency — Inverse Document Frequency (TF-IDF) words representation to the SVM model. Finally, they presented four deep learning models that can notice and classify Arabic hate speech on Twitter into several types.

In [27], the authors were the first who addressed the problem of recognizing speech encouraging religious hatred in the Arabic Twitter. Thus, they were able to detect messages that use provocative sectarian speech to promote hatred and violence against people based on their religious beliefs. They found that a simple Recurrent Neural Network (RNN) architecture with Gated Recurrent Units (GRU) can adequately detect religious hate speech. The used data set is available online at [28]. The authors of [29] presented the foremost publicly-available Levantine Hate Speech and Abusive (L-

HSAB) Twitter dataset. It is intended to be a benchmark dataset for automatic detection of online Levantine harmful contents. The dataset, which is available at [30], includes 5,846 tweets that could be of Normal class, Abusive, or Hate speech.

Considerable work has been investigated for hatred speech detection in the English language. However, rare work has targeted the detection of hate speech in the Arabic language. The majority of the Arabic research targeted web pages and search engines, while a few targeted comments on social networks. In this work, we target the Arabia language and use the data set of [28]. Thus, our constructed models would be mainly compared with [27].

#### IV. PROPOSED METHODOLOGY

The proposed architecture for Arabic hate speech detection is showing in Fig. 2. It includes the subsequent major steps: collection of labelled text document/tweet, text preprocessing, text representation and feature extraction, building of classification models (learning), and Relearning (testing) and classification process.

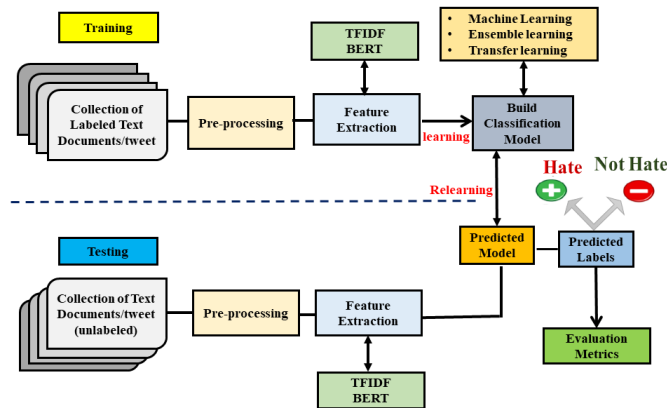


Fig. 2. The Proposed Architecture for Hate Speech Detection Model.

##### A. Data set: Collection of Labelled Text Document/Tweet

In this work we used the data set which was collected by [27] and it is available online at [28]. The data set contains 6164 Arabic tweets and concentrates on the four most typical sacred religions in the Middle East, which are Islam (93.0%), Christianity (3.7%), Judaism (1.6%), and Atheism (0.6%). Originally, the training data set contains 5,569 examples, while the testing data set contains 567 document. The data works for binary classification with two hateful and non hateful classes represented by 1 and 0 respectively. Since data re-sampling is utilized to settle the issue of an imbalanced data set, we performed re-sampling technique. According, the model built with data oversampling is called *Classifier-Over* while the model built with data under sampling is called *Classifier-Under*, where *Classifier* could be any of the six models we used.

##### B. Data Pre-Processing

Text pre-processing includes various techniques that prepare text for further processing. Pre-processing aims to remove the unwanted words from the text, e.g., punctuation, slang, and

stop words. Usually, we have to deal with various preprocessing techniques and a combination of them, including:

1) *Tokenization*: Tokenization is the activity of splitting text into terms, phrases, symbols or additional important elements, called tokens [31]. The obtained elements can be single items (1-gram) or a series of n words (n-gram). Items can be phonemes, syllables, letters, words or even sentences.

2) *Stopwords Removal*: Our work targets Arabic text. Thus, for pre-processing stage, we first remove the non-Arabic text. Every non-Arabic character is replaced with a whitespace character. Moreover, we remove *stopwords*, which appear frequently in the text and are not important for text classification, e.g., مع، أو، في، على، عن، لكن. A list of the most frequently used Arabic stop words is available at [20]. Approximately, 20%–30% of the total words in a record are stopwords, that is, terms that can be removed as they are redundant without any semantic value [32]. The traditional approach for extracting stopwords includes a pre-filled list, containing all words that are semantically irrelevant to a specific language. This technique is a static. On the other hand, the stopwords are recognized online and not specified previously for the dynamic technique. The features are specified based on their importance. Similar to the removal of stopwords, this work eliminates the punctuation and digits from the Arabic text.

3) *Stemming and Lemmatization*: In the Arabic language, various words could be generated from the basic/root word. For example, the words لاعبة، يلعب، ملعب، لاعب، لاعبة are derived from the word لعب. Thus, the stemming operation is applied to reduce the words into their stems. Stemming algorithms can be categorized into three classes: truncation, statistical and mixed techniques. This work conducts Light Stemming for Arabic words to reduce words to their stems. Light stemming withdraws common affixes from words without declining them to their stems. The main idea is that numerous word variants do not have identical meanings although they are developed from the same root. Light stemming aims to improve feature drop while keeping the words' meanings. It removes some specified prefixes and suffixes from the word instead of removing the original root. Lemmatization is a pre-processing approach similar to stemming; the purpose is to decrease the morphological forms of a word to its lemma.

There are many approaches proposed for stemming Arabic words, e.g., light stemming, morphological analysis, statistical-based stemming, and N-grams. Some approaches are language-independent while other approaches are language-dependent. Statistical approaches are language-dependent. Thus, can be tailored for Arabic. Light stemming does not reduce the word into a three-letter stem. However, it just expels the prefixes and suffixes and can achieve good information retrieval without morphological studies.

##### C. Text Representation/ Feature Engineering

The feature selection procedure allows selecting some of the initial feature set, removing the attributes with little predictive capability. For example, wrapper methods in WEKA, execute an investigation over the potential subsets of the initial feature set, assessing the implementation of a classifier over

each one. However, wrapper methods are unusable for large problems. Thus, they are discarded in text classification. On the other hand, filter methods are independent of the classifier with a less computational expense. Filters applied before using the feature selection metric incorporates the removal of infrequent words and overly common words. There are various techniques to convert string data into numerical data such as Bag of words (BoW), Term Frequency — Inverse Document Frequency (TFIDF), Word2Vec, and Bidirectional Encoder Representations from Transformers (BERT). In the following section, some of these techniques will be explained [33][13].

1) *Bag of Words (BoW)*: BoW is a textual representation method suitable for classification models, where the text is viewed as a set of words without considering its syntax or semantics. BoW reveals whether a word is present in the document or not, where the order of the words in the document is insignificant [34]. While constructing the BoW the list of stop words is excluded since they appear frequently with little of useful information. The performance of various ML methods we built utilizing BoW was poor due to the loss of semantic and syntactic information between words. Thus, we used other representation techniques that can handle semantics and syntactic in order to increase performance.

2) *Term Frequency — Inverse Document Frequency (TF-IDF)*: Term Frequency (TF) is a well-known textual representation model which is similar to the BoW technique. However, TF relies on the recurrence of the term in a provided text, while BoW depends on its presence. TF is the frequency of any *term* in a given *document*, which is expressed as given in Equation 1. However, words that are *common* in every document, such as *articles*, *conjunctions*, and *prepositions* rank low because they don't express much to the document. Therefore, we use Inverse Document Frequency (IDF) to reduce the significance of phrases that occur very often in the document collection and improve the importance of phrases that occur infrequently. IDF is constant per corpus and accounts for the ratio of documents that include that specific *term*. It is expressed as given in Equation 2. TF-IDF is a statistical standard to assess how much a phrase is related to a manuscript in a set of documents, i.e., corpus. TF-IDF is computed by multiplying TF by IDF. TF-IDF is regarded as a simple procedure for text classification. Thus, the TF-IDF is developed during model training and then utilized for the test set.

$$TF = \frac{\text{Number of times a term appear in the document}}{\text{Total number of terms in the document}} \quad (1)$$

$$IDF = \text{Log}_{10} \frac{\text{Total number of Documents}}{\text{Number of documents that includes the term}} \quad (2)$$

3) *Word2vec*: Word2vec is a word-embedding technique. It is useful in constructing guidance engines and making sense of sequential data [35]. Word2vec is a prediction-based approach built based on a persistent bag-of-words (CBOW) and a skip-gram (SG). These measures utilize small neural networks (NN) to realize the mapping of words to a point in a vector space. To train the word2vec, the number of the embedding dimensions

is set between 50 and 500 while the length of the context window is set between 5 and 10 [36].

4) *Bidirectional Encoder Representations from Transformers (BERT)*: BERT is a contextualized word representation model founded on a multilayer bi-directional transformer-encoder, where the transformer neural network uses parallel attention layers rather than sequential recurrence [37]. The authors of [37] introduced **BERT** (Bidirectional Encoder Representations from Transformers), where the proposed framework includes two phases: (1) pre-training: the model is prepared on unlabeled data over various pre-training tasks, and (2) fine-tuning: the model is initialized with the pre-trained parameters. Then, all of the parameters are fine-tuned using labeled data from the downstream tasks. Thus, deep bidirectional architectures of BERT allow the same pre-trained model to successfully embark on a broad set of natural language processing tasks. In this work, we used TF-IDF and Bidirectional Encoder Representations from Transformers for the Arabic language called (AraBERT).

#### D. Building of Classification Models (Learning)

We utilized the six ML-based models described next which are Gradient Boosting (GB), K-Nearest Neighbor (K-NN), Logistic Regression (LR), Naive Bayes (NB), Passive Aggressive Classifier (PAC), Support Vector Machine (SVM). Thus, we built the classifiers for these models. The default data partitioning was 80/20 where 80% of the data are used for building the classification model and the remaining 20% are used for model testing. For additional investigation, we built the same models/classifiers for 70/30 data partitioning as well as 90/10. Thus, we have a total of 18 configurations. Moreover, we used transfer learning and build a model based on Ara-BERT and another model called AJGT-BERT. The evaluation of the 20 models/classifiers we built based on different classification metrics is explained in Section VI.

1) *Gradient Boosting*: Boosting algorithm is an ensemble learning algorithm utilizing learning theory [38]. Thus, an group of weak classifiers with low classification accuracy are used to build a strong classifier with higher accuracy. The training procedure of expanding algorithm is incremental, i.e., develops a new classifier in each iteration. Thus, the classifier ranks all instances to evaluate the importance of each instance. Then, the importance of the earlier samples with misclassification are improved. Finally, a stable and better performance classification model is obtained. The gradient boosting (GB) algorithm [38] is an amended algorithm based on the classic boosting algorithm, where it shows better learning ability. Like boosting, GB builds the model with an iterative design, but the model is extended with an optimized loss function. Gradient Boosting is a method drawing awareness for its prediction quickness and accuracy, particularly with extensive and complicated data. Building a GB model start by creating a single leaf rather than a tree or a stump. This leaf symbolizes an starting prediction for the class of all instances. Like AdaBoost, Gradient Boosting build a fixed sized tree based on previous tree's errors where each tree can be larger than a stump. GB scales all the trees by the same amount.

2) *K-Nearest Neighbor (K-NN)*: The k-nearest neighbors (KNN) algorithm is a straightforward, easy-to-implement supervised ML algorithm that is applicable for both classification

and regression [33]. K-NN supposes the likeness between the recent and the known cases. Then, place the recent case into the class that is most identical to the available classes. K-NN algorithm keeps all the available data and categorizes a new instance based on the similarity. K-NN algorithm does not make any hypothesis on underlying data. Thus, it is a non-parametric algorithm [23]. KNN is a lazy learner algorithm because it holds the dataset and it achieves an activity on the dataset at classification, i.e., does not memorize from the training set immediately. However, as the number of independent variables increases the algorithm gets incredibly slower.

3) *Logistic Regression (LR)*: It is comparable to linear regression. However, it expects if a value is True or False rather than predicting a continuous value. Linear regression fits a curve to the data while LR fits an “S” shaped logistic function [23]. Logistic regression can perform on both continuous and discrete data. Thus, its ability to predict the probability and classify new samples makes it a popular ML method, where it is referred to as a probabilistic classifier since it predicts the probability of an output. Usually, logistic regression is used for classification. In linear regression, we fit the line between the data using “least squares”. However, the concept of “residual” does not apply to LR where the concept of “maximum likelihood” is rather used.

4) *Naive Bayes (NB)*: Naive Bayes classifier relies on Bayes Theorem, which works on conditional probability. The conditional probability is the likelihood that something will happen, given that something else has already occurred [39] [40]. The formula for calculating the conditional probability is given in Equation 3, where  $H$  is the hypothesis and  $E$  is the evidence.

$$P(H|E) = \frac{P(E|H) \times P(H)}{P(E)} \quad (3)$$

For a set of labeled training data, NB evaluates different model parameters, e.g., the likelihood of each class label to appear. Then, predict the class for any given test data based on its probability to be assigned for different classes. The maximum probability determines the predicted class [23].

5) *Passive Aggressive Classifier (PAC)*: This is an online ML algorithm, where it responds as passive for correct classifications and as aggressive for any miscalculation. In PAC we train a system incrementally by providing it samples sequentially, i.e., individually or in small groups called mini batches [33] [41]. The primary principle of this algorithm is that it notices data, learns from the data, and discards it without the need of storing the data. However, in batch learning the entire training dataset is used at once. Thus, PAC is suitable for systems that acquire data in a steady stream such as news and social media [42]. When the prediction is correct, we keep the model without any changes since the data in the example was not enough to change the model. Thus, it is called Passive. However, for incorrect prediction we introduce some changes to the model that could correct it. Thus, it is called Aggressive. PAC algorithm proved its effectiveness for online learning to solve various real-world problems [43].

6) *Support Vector Machine (SVM)*: SVM is a supervised classifier. For a set of labeled training data, SVM realizes

a hyperplane that distinctly classifies the data points while maximizing the margin between the data instances and the hyperplane itself [13] [44]. Then, the class of test data is determined based on the realized hyperplane [42].

7) *AraBERT*: Based on BERT [37], the authors of [45] presented **AraBERT** (transformer-based Model for Arabic Language Understanding), where they pre-trained BERT, especially for the Arabic language aiming to achieve the same success as BERT. The authors of [45] used the original configuration of BERT which has 12 encoder blocks, 768 hidden dimensions, 12 attention heads, 512 maximum sequence length, and a total of 110M parameters. Then, to better fit the Arabic language, they introduced additional preprocessing before model’s pre-training. To avoid information loss, they maintained words with Latin characters so they can cite named entities and scientific phrases in their original language. Thus, after eliminating duplicate sentences, the final size of the pre-training data sets was 70 million sentences.

8) *AJGT-BERT*: The authors of [46] introduced an Arabic tweets corpus documented in Jordanian dialect and Modern Standard Arabic (MSA) annotated for sentiment analyses. The generated AJGT corpus consists of 1,800 tweets with 900 classified as positives and the remaining 900 are negatives. The Arabic Jordanian General Tweets (AJGT) data sets is publicly available online at [46].

## V. EXPERIMENTAL ANALYSIS

A  $2 \times 2$  matrix, which is called a confusion matrix, is created to visually illustrate the performance of a binary supervised learning problem. Table I shows the confusion matrix for Arabic hate speech detection. It includes four classes, which are true positive, true negative, false positive, and false negative. In this work, **True Positive** (TP) indicates that the comment is actually hate speech and correctly classified as hate speech. **True Negative** (TN) indicates that the comment is non-hate speech and correctly classified as non-hate speech. **False Positive** (FP) means that the comment is actually non-hate speech but incorrectly classified as hate speech, and **False Negative** (FN) describes the comment that is actually hate speech but incorrectly classified as non-hate speech. For any classification model, we aim to maximize the value of TP and TN and minimize the value of FP and FN.

TABLE I. CONFUSION MATRIX OF ARABIC HATE SPEECH DETECTION

	Actual Hate Speech	Actual Non-Hate Speech
Predicted Hate Speech	True Positive(TP)	False Positive(FP)
Predicted Non-Hate Speech	False Negative (FN)	True Negative(TN)

### A. Implementation Environment

To accomplish all investigations in this work, we utilized a PC with the following details: Intel R © Core(TM) i7-6850 K processor with 8 GB RAM and 3.360 GHz frequency. Regarding the software, we have used Python 3.8.0 programming with Anaconda [Jupyter notebook] for ML and Colab for transfer learning models. We used various libraries such as NumPy, Pandas, Sci-kit-learn TensorFlow, and Keras.

## B. Evaluation Metrics

As given in Equation 4, **accuracy** denotes the number of rightly classified data samples over the total number of data samples. However, for an unbalanced dataset, where positive and negative classes have a different number of instances, the accuracy is not suitable to evaluate the model. **Precision** (positive predictive value) as defined in Equation 5, should be 1 for a perfect classifier while the value of FP is zero. **Recall** which is known as sensitivity or true positive rate is defined as given in Equation 6. For a perfect classifier, recall should be 1 while the value of FN is zero. For an ideal classifier, both precision and recall are 1. **F1-score** is a metric that depends on both precision and recall and is defined as given in Equation 7. F1-score becomes 1 only when precision and recall are both 1. So, F1-score is the harmonic mean of precision and recall and it is a better measure than accuracy [41] [44].

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (4)$$

$$Precision = \frac{TP}{TP + FP} \quad (5)$$

$$Recall/Sensitivity = \frac{TP}{TP + FN} \quad (6)$$

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (7)$$

## VI. DISCUSSION AND RESULTS

In this section, we will explain the result of various machine learning models which we have used in this study. We have designed different models to detect and classify religious Arabic hate speech based on various methods, e.g., data partitioning, re-sampling and transfer learning. For that, we have divided our data (train/test) for three scenarios. The data is partitioned into (70/30), (80/20), and (90/10). For each partitioning, we use the original data sets in addition to the oversampling and under-sampling techniques. The best-obtained classification performance in partitioning scenarios was for (80/20), where a detailed explanation is given in Section VI-A. Then, Section VI-B explains the classification performance for 70/30 data partitioning while Section VI-C is dedicated to 90/10 data partitioning. The proposed models were evaluated on a testing data set related to religious hate speech in Arabic text [28]. In order to enhance the performance of the classifiers that we build for six ML algorithms, we have extended this implementation to include transfer learning methods. The transfer learning models called Ara-BERT and AJGT-BERT. The comparison of the performances of all models have been done in terms of accuracy, recall, precision, and F1 score using Arabic hate speech data set.

## A. Hate Speech Detection for (80/20) Data Partitioning

Fig. 3 shows the various obtained classification metrics based on various models, i.e., GB, K-NN, LR, NB, PAC, SVM, Ara-BERT, and BERT-AJGT. Clearly, the Ara-BERT model achieves the best classification metrics followed by AJGT-BERT while the KNN classifier has the lowest metrics. A detailed explanation for each metric is given next.

The obtained **precision** for the original data without re-sampling based on 8 classifiers is shown in Fig. 3. Both PAC and SVC has a precision of 75%. Moreover, the transfer-based classifiers, i.e., **Ara-BERT** and AJGT-BERT, have the highest precision with 79% and 78%, respectively. KNN classifier has the lowest precision of 69%. The obtained **Recall** based on 8 classifiers without re-sampling are very similar to the obtained precision. The transfer-based classifiers, i.e., **Ara-BERT** and AJGT-BERT, have the highest Recall with 79% and 78%, respectively, while KNN classifier has the lowest precision of 69%. Moreover, both PAC and SVC have a Recall of 75%. As given in Equation 7, **F1-Score** depends on both precision and recall. Regarding the obtained F1-Score for the various classifiers, the transfer-based classifiers, i.e., **Ara-BERT** and AJGT-BERT, have the highest F1-Score with 79% and 78%, respectively, while KNN classifier has the lowest F1-Score of 69%. When evaluating the **accuracy** we notice that it is very close to F1-Score of the various classifiers. Transfer-based classifiers are the highest while KNN classifier has the lowest accuracy, i.e., the accuracy of **Ara-BERT** model is 79%. Next, we are going to explain the classification metrics with data over-sampling and under-sampling.

## Hate Speech Detection for (80/20) Data Partitioning with Oversampling

With oversampling, we apply oversampling technique by increase the minority of samples to be same to majority like 2196 to 3650. Then, we used 80% of the data to construct the classification model and the rest 20% for testing the model/classifier. The obtained Precision, Recall, F1-Score and Accuracy of the various classifiers are shown in Fig. 3 where the name of a specific classifier is indicated as *Classifier-Over*. The various metrics for the Gradient Boosting and K-Nearest Neighbor (K-NN) classifiers with data oversampling remains the same as the original data set. However, with over-sampling the LR and NB classifiers have a 1% to 2% improvement of Precision and F-Score. Data oversampling reduces the various metrics of PAC by 2%. However, SVM based classifier is unaffected by data oversampling. Transfer-based classifiers are unaffected by oversampling. Thus, **Ara-BERT** has a value of 79% for Precision, Recall, F1-Score and Accuracy.

## Hate Speech Detection for (80/20) Data Partitioning with Under-Sampling

With under-sampling, we decreased the number of normal data from 3650 to 2196 to be equal to the hate speech. Then, we employed 80% of the data to construct the model and the remaining 20% for testing the model/classifier. The obtained Precision, Recall, F1-Score and Accuracy of the various classifiers are shown in Fig. 3 where the name of a specific classifier is indicated as *Classifier-Under-sampling*. With under-sampling, the various metrics of GB classifier are



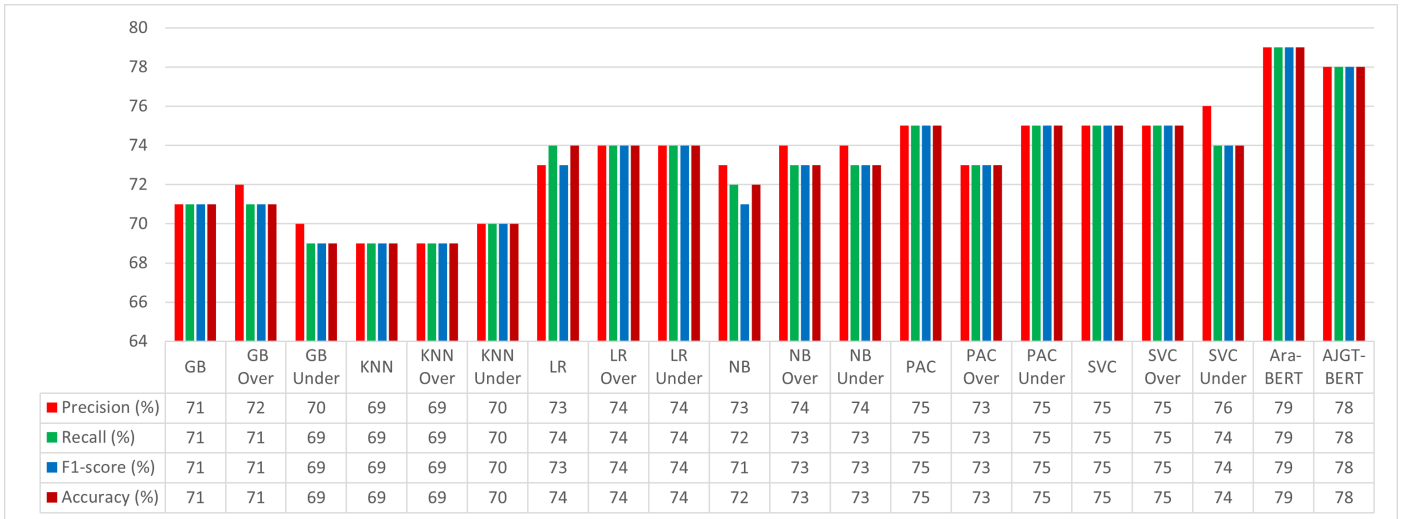


Fig. 3. Various Classification Metrics for 80/20 Data Partition with Different Models.

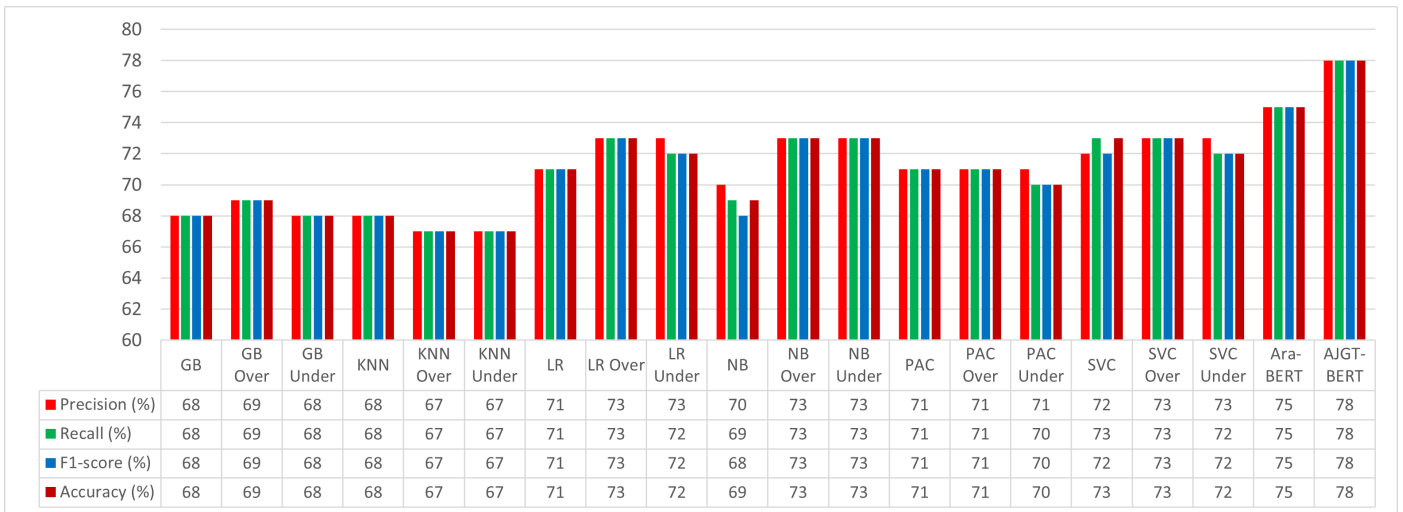


Fig. 4. Various Classification Metrics for 70/30 Data Partition with Different Models.

reduced by 1% while they are increased by 1% for the KNN classifier. Similarly, the Precision and F1-Score of the LR classifier are enhanced by 1% while its recall and accuracy are unchanged. All metrics of the NB classifier are increased by 1% for under-sampling while they remain the same for the PAC classifier. The precision of SVC is increased by 1% to reach 76% while its recall, F1-Score and Accuracy are reduced by 1% to become 74% for all of them. Transfer-based classifiers are unaffected by under-sampling. Thus, **Ara-BERT** has a value of **79%** for Precision, Recall, F1-Score and Accuracy while the AJGT-BERT classifier has a value of 78% for the same metrics.

Based on the shown result for 80/20 data partitioning, the Ara-BERT models archives the best evaluation metrics with 79% for the precision, recall, F1-Score and accuracy. Data re-sampling introduced a 1% to 2% improvement where such insignificant gain is due to the original distribution of the training data.

### B. Hate Speech Detection for (70/30) Data Partitioning

Fig. 4 shows the various classification metrics for the different models with 70/30 data partitioning. AJGT-BERT classifier has the best metrics of 78% (for precision, recall, F1-Score and accuracy) while Ara-BERT classifier has a value of 75%. There are various classifiers that achieves 73% for all classification metrics including, LR-Over, NB-Over, NB-Under, and SVC-Over. We notice that sometimes data re-sampling introduces a minor improvement of 1% to 2% in few classifiers.

### C. Hate Speech Detection for (90/10) Data Partitioning

Fig. 5 shows the various classification metrics for the different models with 90/10 data partitioning. AJGT-BERT and SVC have the highest performance with 78% (for precision, recall, F1-Score and accuracy) followed by SVC with oversampling (SVC-Over) with 77% performance. Ara-BERT, SVC-Over,

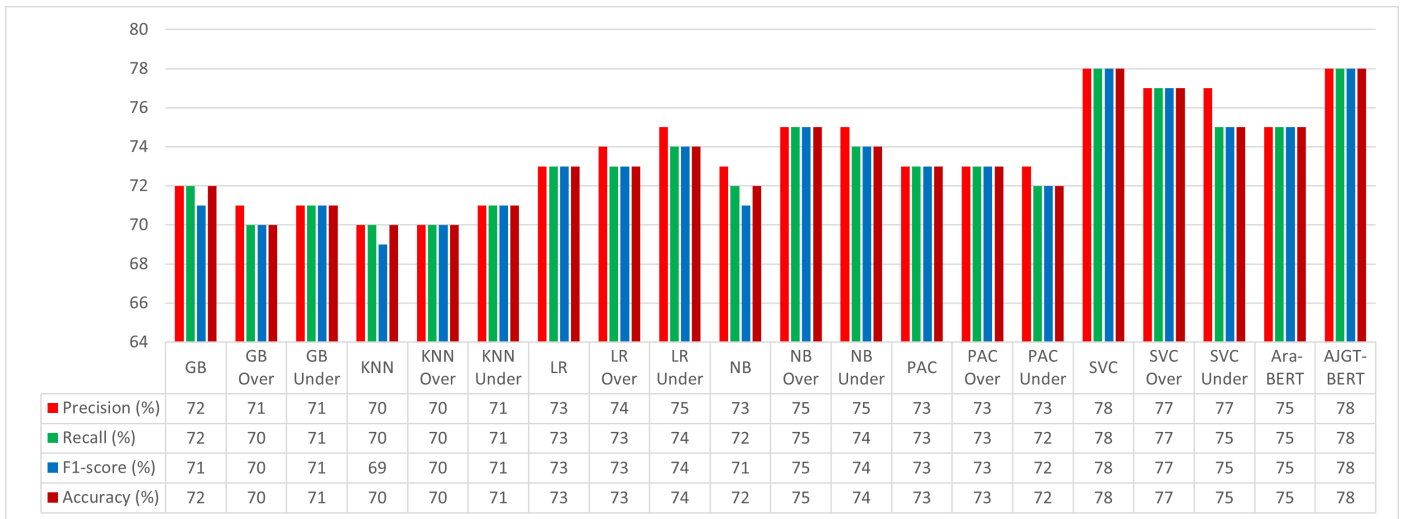


Fig. 5. Various Classification Metrics for 90/10 Data Partition with Different Models.

TABLE II. COMPARISON OF CLASSIFICATION RESULTS OVER [28] DATASET

	Precision	Recall	F1-Score	Accuracy
Related work [27]	76	78	77	79
Ara-BERT with 70/30	75	75	75	75
AJGT-BERT with 70/30	<b>78</b>	<b>78</b>	<b>78</b>	<b>78</b>
Ara-BERT with 80/20	<b>79</b>	<b>79</b>	<b>79</b>	<b>79</b>
AJGT-BERT with 80/20	<b>78</b>	<b>78</b>	<b>78</b>	<b>78</b>
Ara-BERT with 90/10	75	75	75	75
AJGT-BERT with 90/10	<b>78</b>	<b>78</b>	<b>78</b>	<b>78</b>
SVC with 90/10	<b>78</b>	<b>78</b>	<b>78</b>	<b>78</b>

and NB-Over classifiers have 75% performance while NB-Under and LR-Under are 74%.

Table II shows a comparison of classification results over [28] where we used their dataset. The table compares various models we build with the related work [27]. Clearly, we were able to build models similar to or better than the related work. In most cases, transfer learning based models have the highest precision, recall, F1-Score, and accuracy, while other ML models have a very similar metrics

## VII. CONCLUSION

Hate speech is one of the major problems at this time, especially with the increasing number of users on social media. At the same time, an increasing number of crimes became a serious concern that threatens the cohesiveness and structure of civilian societies. Therefore, this work presents an efficient framework to detect Arabic hate speech based on the content of social networks. We utilize various ML and DL models to perform an efficient classification of users' comments. Based on the content of this work, the classes are hate speech or normal. The proposed framework has six ML algorithms and two DL, which are Gradient Boosting, K-Nearest Neighbor, Logistic Regression, Naive Bayes, Passive Aggressive Classifier, Support Vector Machine, Ara-BERT, and BERT-AJGT. For wider investigation, we utilized various scenarios of data partitioning, re-sampling techniques, and transfer learning. We were able to successfully have various classification models

with better results in terms of precision, recall, F1-Score, and accuracy compared to the most relevant related work. For future work, we aim to create huge and benchmark data sets. Moreover, working with the mixed language problem, multi-model and data augmentation can be interesting topics for future work on this topic, especially for the Arabic language. In future work, the classification can be expanded to cover many classes such as racism, misogyny, religious discrimination and so on.

## REFERENCES

- [1] H. Butt, M. R. Raza, M. J. Ramzan, M. J. Ali, and M. Haris, "Attention-based CNN-RNN Arabic text recognition from natural scene images," *Forecasting*, vol. 3, no. 3, pp. 520–540, 2021.
- [2] R. Duwairi, A. Hayajneh, and M. Quwaider, "A deep learning framework for automatic detection of hate speech embedded in arabic tweets," *Arabian Journal for Science and Engineering*, vol. 46, no. 4, pp. 4001–4014, 2021.
- [3] I. Guellil, H. Saädane, F. Azouaou, B. Gueni, and D. Nouvel, "Arabic natural language processing: An overview," *Journal of King Saud University-Computer and Information Sciences*, vol. 33, no. 5, pp. 497–507, 2021.
- [4] K. Darwish, W. Magdy, and A. Mourad, "Language processing for arabic microblog retrieval," in *21st ACM international conference on Information and knowledge management*, 2012, pp. 2427–2430.
- [5] J. Chen, Y. Chen, Y. He, Y. Xu, S. Zhao, and Y. Zhang, "A classified feature representation three-way decision model for sentiment analysis," *Applied Intelligence*, vol. 52, no. 7, pp. 7995–8007, 2022.
- [6] G. Kovács, P. Alonso, and R. Saini, "Challenges of hate speech detection in social media," *SN Computer Science*, vol. 2, no. 2, pp. 1–15, 2021.
- [7] A. Matamoros-Fernández and J. Farkas, "Racism, hate speech, and social media: A systematic review and critique," *Television & New Media*, vol. 22, no. 2, pp. 205–224, 2021.
- [8] K. Barker and O. Jurasz, "Online misogyny as a hate crime:# timesup," in *Misogyny as Hate Crime*. Routledge, 2021, pp. 79–98.
- [9] A. Al-Hassan and H. Al-Dossari, "Detection of hate speech in social networks: a survey on multilingual corpus," in *6th international conference on computer science and information technology*, vol. 10, 2019.
- [10] D. Sultan, A. Suliman, A. Toktarova, B. Omarov, S. Mamikov, and G. Beissenova, "Cyberbullying Detection and Prevention: Data Mining in Social Media," in *International Conference on Cloud Computing, Data Science & Engineering*. IEEE, 2021, pp. 338–342.

- [11] S. U. Maheswari and S. Dhenakaran, "Analysis of Approaches for Irony Detection in Tweets for Online Products," in *Innovations in Computational Intelligence and Computer Vision*. Springer, 2022, pp. 141–151.
- [12] I. Kwok and Y. Wang, "Locate the hate: Detecting tweets against blacks," in *Twenty-seventh AAAI conference on artificial intelligence*, 2013.
- [13] A. Y. Muaad, H. J. Davanagere, D. Guru, J. Benifa, C. Chola, H. Al-Salman, A. H. Gumaei, and M. A. Al-antari, "Arabic document classification: Performance investigation of preprocessing and representation techniques," *Mathematical Problems in Engineering*, vol. 2022, 2022.
- [14] I. Aljarah, M. Habib, N. Hijazi, H. Faris, R. Qaddoura, B. Hammo, M. Abushariah, and M. Alfawareh, "Intelligent detection of hate speech in arabic social network: A machine learning approach," *Journal of Information Science*, vol. 47, no. 4, pp. 483–501, 2021.
- [15] "Jordanian Ministry of Justice," last accessed September 14, 2022. [Online]. Available: <http://www.moj.gov.jo/EchoBusV3.0/SystemAssets/5d38ea27-5819-443e-a380-b65c7e1f5b56.pdf>
- [16] M. Masadeh, A. Masadeh, O. Alshorman, F. Khasawneh, and M. Masadeh, "An efficient machine learning-based covid-19 identification utilizing chest x-ray images," *IAES International Journal of Artificial Intelligence*, pp. 356–366, 2022.
- [17] M. Masadeh, O. Hasan, and S. Tahar, "Machine-learning-based self-tunable design of approximate computing," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 29, no. 4, pp. 800–813, 2021.
- [18] —, "Machine learning-based self-compensating approximate computing," in *2020 IEEE International Systems Conference (SysCon)*. IEEE, pp. 1–6.
- [19] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.
- [20] R. NL. Stopword lists. [Online]. Available: <https://www.ranks.nl/stopwords/arabic>
- [21] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He, "A Comprehensive Survey on Transfer Learning," *Proceedings of the IEEE*, vol. 109, no. 1, pp. 43–76, 2021.
- [22] R. Mohammed, J. Rawashdeh, and M. Abdullah, "Machine learning with oversampling and undersampling techniques: Overview study and experimental results," in *11th International Conference on Information and Communication Systems (ICICS)*, 2020, pp. 243–248.
- [23] M. K. A. Aljero and N. Dimililer, "A Novel Stacked Ensemble for Hate Speech Recognition," *Applied Sciences*, vol. 11, no. 24, p. 11684, 2021.
- [24] S. MacAvaney, H.-R. Yao, E. Yang, K. Russell, N. Goharian, and O. Frieder, "Hate speech detection: Challenges and solutions," *PLOS ONE*, vol. 14, no. 8, pp. 1–16, 08 2019.
- [25] T. Putri, S. Sriadhi, R. Sari, R. Rahmadani, and H. Hutahaean, "A comparison of classification algorithms for hate speech detection," in *IOP Conference Series: Materials Science and Engineering*, vol. 830, no. 3. IOP Publishing, 2020, p. 032006.
- [26] A. Al-Hassan and H. Al-Dossari, "Detection of hate speech in arabic tweets using deep learning," *Multimedia Systems*, pp. 1–12, 2021.
- [27] N. Albadi, M. Kurdi, and S. Mishra, "Are they our brothers? analysis and detection of religious hate speech in the arabic twittersphere," in *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE, 2018, pp. 69–76.
- [28] "Religious Hate Speech Detection for Arabic Tweets," last accessed September 14, 2022. [Online]. Available: [https://github.com/nuhaalbadil/Arabic\\_hatespeech](https://github.com/nuhaalbadil/Arabic_hatespeech)
- [29] H. Mulki, H. Haddad, C. B. Ali, and H. Alshabani, "L-HSAB: A Levantine Twitter Dataset for Hate Speech and Abusive Language," in *Proceedings of the third workshop on abusive language online*, 2019, pp. 111–118.
- [30] Hala-Mulki. First-arabic-levantine-hatespeech-dataset. [Online]. Available: <https://github.com/Hala-Mulki/L-HSAB-First-Arabic-Levantine-HateSpeech-Dataset/blob/master/Dataset/L-HSAB>
- [31] M. O. Hegazi, Y. Al-Dossari, A. Al-Yahy, A. Al-Sumari, and A. Hilal, "Preprocessing Arabic text on social media," *Heliyon*, vol. 7, no. 2, p. e06191, 2021.
- [32] T. Kanan, B. Hawashin, S. Alzubi, E. Almaita, A. Alkhatib, K. A. Maria, and M. Elbes, "Improving arabic text classification using p-stemmer," *Recent Advances in Computer Science and Communications (Formerly: Recent Patents on Computer Science)*, vol. 15, no. 3, pp. 404–411, 2022.
- [33] A. Y. Muaad, H. Jayappa Davanagere, J. Benifa, A. Alabrah, M. A. Naji Saif, D. Pushpa, M. A. Al-Antari, and T. M. Alfakih, "Artificial intelligence-based approach for misogyny and sarcasm detection from arabic texts," *Computational Intelligence and Neuroscience*, vol. 2022, 2022.
- [34] J. Salminen, M. Hopf, S. A. Chowdhury, S.-g. Jung, H. Almerexhi, and B. J. Jansen, "Developing an online hate classifier for multiple social media platforms," *Human-centric Computing and Information Sciences*, vol. 10, no. 1, pp. 1–34, 2020.
- [35] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.
- [36] F. K. Khattak, S. Jeblee, C. Pou-Prom, M. Abdalla, C. Meaney, and F. Rudzicz, "A survey of word embeddings for clinical text," *Journal of Biomedical Informatics*, vol. 100, p. 100057, 2019.
- [37] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.
- [38] L. Mason, J. Baxter, P. Bartlett, and M. Frean, "Boosting algorithms as gradient descent," *Advances in neural information processing systems*, vol. 12, 1999.
- [39] M. Masadeh, A. Aoun, O. Hasan, and S. Tahar, "Decision tree-based adaptive approximate accelerators for enhanced quality," in *International Systems Conference (SysCon)*. IEEE, 2020, pp. 1–5.
- [40] A. Elouardighi, M. Maghfour, H. Hammia, and F.-z. Aazi, "A machine learning approach for sentiment analysis in the standard or dialectal arabic facebook comments," in *3rd International Conference of Cloud Computing Technologies and Applications (CloudTech)*, 2017, pp. 1–8.
- [41] A. Y. Muaad, G. H. Kumar, J. Hanumanthappa, J. B. Benifa, M. N. Mourya, C. Chola, M. Pramodha, and R. Bhairava, "An effective approach for arabic document classification using machine learning," *Global Transitions Proceedings*, vol. 3, no. 1, pp. 267–271, 2022.
- [42] K. Nagashri and J. Sangeetha, "Fake news detection using passive-aggressive classifier and other machine learning algorithms," in *Advances in Computing and Network Communications*. Springer, 2021, pp. 221–233.
- [43] J. Lu, P. Zhao, and S. C. Hoi, "Online Passive-Aggressive Active Learning," *Machine Learning*, vol. 103, no. 2, pp. 141–183, 2016.
- [44] A. Y. Muaad, H. J. Davanagere, M. A. Al-antari, J. B. Benifa, and C. Chola, "Ai-based misogyny detection from arabic levantine twitter tweets," in *Computer Sciences & Mathematics Forum*, vol. 2, no. 1. MDPI, 2021, p. 15.
- [45] W. Antoun, F. Baly, and H. M. Hajj, "AraBERT: Transformer-based Model for Arabic Language Understanding," *CoRR*, vol. abs/2003.00104, 2020.
- [46] K. M. Alomari, H. M. ElSherif, and K. Shaalan, "Arabic tweets sentimental analysis using machine learning," in *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*. Springer, 2017, pp. 602–610.

# Modeling Multioutput Response Uses Ridge Regression and MLP Neural Network with Tuning Hyperparameter through Cross Validation

Waego Hadi Nugroho<sup>1</sup>, Samingun Handoyo<sup>2</sup>, Hsing-Chuan Hsieh<sup>3</sup>, Yusnita Julyarni Akri<sup>4</sup>, Zuraidah<sup>5</sup>, Donna DwinitaAdelia<sup>6</sup>

Department of Statistics, Brawijaya University, Malang 65145, Indonesia<sup>1,2</sup>

Department of Electrical Eng. And Computer Sci.-IGP, National Yang Ming Chiao Tung University<sup>2</sup>  
Hsinchu 30010, Taiwan<sup>2</sup>

Institute of Statistics, National Yang Ming Chiao Tung University, Hsinchu 30010, Taiwan<sup>3</sup>

Department of Midwifery, Tribuana Tunggadewi University, Malang 65144, Indonesia<sup>4</sup>

Study Program of Islamic Banking, State Islamic Institute of Kediri, Kediri 64127, Indonesia<sup>5</sup>

Department of Midwifery, Wira Husada Nusantara Health Polytechnic, Malang 65144, Indonesia<sup>6</sup>

**Abstract**—The multiple regression model is very popular among researchers in both field of social and science because it is easy to interpret and have a well-established theoretical framework. However, the multioutput multiple regression model is actually widely applied in the engineering field because in the industrial world there are many systems with multiple outputs. The ridge regression model and the Multi-Layer Perceptron (MLP) neural network model are representations of the predictive linear regression model and predictive non-linear regression model that are widely applied in the world of practice. This study aims to build multi-output models of a ridge regression model and an MLP neural network whose hyperparameters are determined by a grid search algorithm through the cross-validation method. The hyperparameter that produces the smallest RMSE value in the validation data is chosen as the hyperparameter to train both models on the training data. The hyperparameter in question is a combination of learning algorithms and alpha values (ridge regression), a combination of the number of hidden nodes and gamma values (MLP neural network). In the ridge regression model for alpha in the range between 0.1 and 0.7, the smallest RMSE is obtained for all learning algorithms used. While the MLP neural network model specifically obtained a combination of the number of nodes = 18 and gamma = 0.1 which produces the smallest RMSE. The ridge regression model with selected hyperparameters has better performance (in the RMSE and R2 value) than the MLP neural network model with selected hyperparameters, both on training and testing data.

**Keywords**—Filter approach; hyperparameter tuning; multi-response; neural network; ridge regression

## I. INTRODUCTION

The health of the mother during pregnancy and the condition of the baby at birth greatly affect the health and intelligence of the younger generation which is the continuation of the sustainability of a nation. Several factors, including the condition of pregnant women, food intake of pregnant women, and health conditions of the family environment affect the condition of the baby at birth, such as

stunting events [1]. The model for predicting the occurrence of a class category (stunting or not stunting) is called a classification model. Comparison of the performance of binary classification models, among others, was carried out by Widodo and Handoyo [2] comparing logistic regression and support vector machine, Handoyo et al [3] comparing logistic regression and Linear Discriminant, while Nugroho et al [4] comparing logistic regression and decision tree on the multiclass label response.

If the response variable has a numerical scale such as the length of the baby at birth [5], the predictive model is called a regression model. Santosa et al [6] used a partial least square approach to explain the effect of factors on maternal and child conditions on stunting where this factor is a latent variable. On the other hand, Sajjad et al [7] used multi-output modeling of response variables. Heating and cooling loads with predictor variables were factors related to the layout of a building. Multi-output response variables derived from the condition of the baby at birth (latent variables) consisting of several numerical indicators are very possible and also a challenge when building a model based on a multi-output system.

Regression modeling using machine learning methods has been applied in various fields, including industrial product design by Turetsky et al [8], wind speed prediction by Barhmi et al [9], prediction of imported soybean prices in Indonesia by Handoyo and Chen [10], and also prediction of beef and chicken prices by Handoyo et al [11]. In general, a model that is free from overfitting problems will have satisfactory performance. The ridge regression model is a multiple regression model which is given a penalty of l2 norm [12]. The regularization technique on the neural network is done by adding a penalty l2 norm to the loss function as an attempt to overcome the overfitting problem [13]. However, tuning hyperparameters on ridge regression and neural networks with regularization is generally done by trial and error. In order for these models to have an optimal combination of hyperparameters (producing the best performance), Tso et al

[14], and also Belete and Huchaiah [15] used the k-folds cross-validation method for hyperparameter tuning.

This study aims to build a multioutput model of ridge regression and MLP neural network on the survey dataset with the predictor and response variables derived from the latent variables. The selection of predictor variables that are free from multicollinearity elements is carried out using the filter approach method. In the ridge regression model, the learning algorithm and alpha values are tuned, while the MLP neural network model is carried out to tune the nodes number in the hidden layer and the gamma value using the grid search method. Evaluation of model performance with RMSE and R2 is carried out on both training and testing data.

## II. RELATED WORKS

Broadly speaking, there are 2 types of modeling in machine learning, namely supervised learning (predictive modeling) and unsupervised learning (descriptive modeling). An unsupervised learning model is characterized by the dataset used in the model building that does not contain a response variable [16]. The response variable measurement unit scale has a critical role, namely if the response variable on a numerical scale will lead to regression modeling, whereas if the response variable is on a categorical scale it will lead to a classification model. Modeling the dengue fever status of a village [17] and modeling the baby's weight status at birth [18] are examples of classification modeling. Regression modeling generally aims to determine the magnitude of the influence of the predictor variable on the numerical response variable and also predicts an unknown value of the response variable based on the values of the predictor variables of a certain instance [19-21]. The research above only involves a single response variable and there has also been no effort to produce a model that is free from overfitting problems.

Often researchers do not pay attention to the unit of measure for each variable contained in the dataset where the action will lead to the incorrect model construction. A commensurate nature of all variables involved in modeling must be maintained so that the arithmetic operations on all formulas used can be guaranteed validity [22]. In addition, correlations between predictor variables should be avoided in order to produce a model that has a low bias value. The selection of predictor variables that are independent of each other can be done before the process of building a model, namely the filter approach method [23]. The filter approach method will greatly reduce the computational cost of complex models involving many parameters [24]. The advantage of the filter approach method is that it reduces the number of predictor variables and still maintains the predictor variables in their original form.

Decision-making in the real world must take into account many factors related to the system being studied. A multi-output model can be a classification or a regression model which if it is given an input, can predict unknown multi-output simultaneously [25]. Assessment of product quality in the food, beverage, and fragrance industries uses a lot of semantic odor perception descriptors. Li et al [26] designed an odor perception descriptor selection mechanism based on a multi-output machine learning model including multiple regression

and neural network to find the main odor perception descriptors. Shams et al [27] compared the performance of Multiple Linear Regression and MLP neural networks to predict SO2 concentration in the air of Tehran. The predictor variables used include meteorological parameters, urban traffic data, urban green open space information, and selected time parameters, while the response variable is the daily concentration of SO2. The MLP model has a better performance than the regression model. Siavash et al [28] predict turbine performance using multiple linear regression and a neural network considering as many as 4 channel opening angles as response variables. The performance of the neural network model is more satisfactory than the multiple regression model. The performance comparison between the regression model and the MLP in the above study did not involve tuning the hyperparameters of both models.

The ridge regression model is widely used in practice because of its ease of interpretation, use, and strong theoretical guarantees. In many cases, the model hyperparameter is tuned by using cross-validation, but when the spectrum of the covariate matrix is almost flat and the observations in the observed model are not too high then cross-validation will be detrimental [29]. Meanwhile, van de Wiel et al [30] proposed fast hyperparameter tuning, and Meanti et al [31] proposed Efficient Hyperparameter tuning in the kernel of ridge regression based on cross-validation of data. Tuning hyperparameters on a neural network model using cross-validation data, among others, was carried out by Blume et al [32], and also by Linder et al [33]. Although there is controversy over the advantages and disadvantages of applying the cross-validation method to set up model hyperparameters, this method is systematic and fair.

## III. PROPOSED METHOD

A model is called a simple model if the model only involves a few predictor variables and the relationship between predictor variables is linear. The most sought-after models are those that are simple and have high performance. Variable selection is needed to avoid multicollinearity between predictor variables and also to reduce the number of predictor variables.

### A. Variable Selection and Data Formatting

In the dataset, each variable is related to its respective units of measure, giving rise to very diverse units of measure. The difference in the unit of measure for each of these variables must be handled in order to meet the rules in arithmetic operations. All variables before being analyzed must have an equivalent unit of measure (commensurate measure). The min-max transformation given to eq.(1) is a simple way to satisfy the commensurate measures of each variable [34].

$$N_i = \frac{P_i - P_{min}}{P_{max} - P_{min}} \quad (1)$$

Where  $N_i$  is the normalized value of the i-th instance,  $P_i$  is the observed value of the i-th instance, and  $P_{min}, P_{max}$  are respectively the minimum and maximum value of the predictor variable P. The eq. (1) will be used to transform all values of the predictor variable P into the range of [0,1] and without a unit of measures.

Variable selection with the filter approach is computationally inexpensive because the selection process does not involve the prospective model to be built. The selection of variables is only based on the level of dependence between two variables. The measuring scale of two variables evaluated for dependence lead to a kind of statistical test, namely the dependence between two categorical variables is evaluated by a chi-square test through a contingency table, and the dependence between numerical and categorical variables is evaluated by a one-way ANOVA test, and the dependence between 2 numerical variables is evaluated by correlation test [35]. The Pearson correlation formula given i.e. Eq. (2) measures a degree of dependence between two numerical variables.

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}} \quad (2)$$

Where r represents a level of dependency between 2 numerical variables namely the x and y variables. The coefficient correlation r has a value in the range between -1 and 1. The value of r = 0 means that there is perfectly no dependency between 2 variables, while the value of r = abs(1) i.e. absolute 1 means that there is a perfect dependence between two variables. To make a simple task in evaluating dependency between two numerical variables, the value of threshold = 0.7 is set. If the value of r is less than absolute 0.7 then the two variables are declared to have no dependency, for the opposite condition means that the two variables have a dependency and as a result one of these variables must be dropped from the dataset [36].

Modeling in machine learning always provides the out-sample data, which is a subset of data obtained by splitting the dataset that separates from the data to build the model. Out-sample data is used to test the model's performance or often referred to as the testing data.

Fig. 1 presents the splitting of the dataset into training and testing parts, also into sub-training, and validation data [37]. In Fig. 1, Initially, the dataset was randomly divided into the training subset (80%) and the testing subset (20%). Furthermore, the training subset is divided randomly into k-fold which are used to form the sub-training and validation data. In this process, k pairs of sub-training and validation data were obtained. For example, if the fold 1 is as the validation data, the other k-1 folds are as the sub-training data, if the fold 2 is as the validation data, the other k-1 folds are as the sub-training data and so on. Model candidates are trained on all sub-training data with each candidate hyperparameter and the model's performance is evaluated on the corresponding validation data. The grid search method is a way to find the model's hyperparameters that give the best average performance on the validation data.

Training					Testing
fold_1	fold_2	fold_3	...	fold_k	
val_1	sub training				
...	...	...	...	...	
sub training					val_k

Fig. 1. The Formatting of the Training Data into k-fold Cross Validation.

### B. Multioutput Multiple Regression and Ridg Regression

In multiple linear regression, if there is more than 1 response variable, it will lead to multi-response modeling which in machine learning is better known as multi-output regression modeling. A simple multi-output regression modeling diagram is given in Fig. 2 as the following.

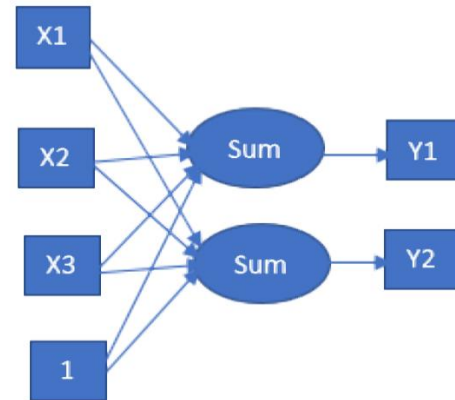


Fig. 2. The Multioutput Multiple Regression Diagram.

In Fig. 2, it is illustrated that there are three predictor variables, namely X1, X2, and X3 as inputs for a system that performs summation operation. This system produces two outputs, namely, Y1 and Y2. In addition, the input system also has a bias of 1. The diagram when expressed in the form of a mathematical formula is as follows:

$$Y_1 = b_1 + w_{11}X_1 + w_{12}X_2 + w_{13}X_3 \quad (3)$$

$$Y_2 = b_2 + w_{21}X_1 + w_{22}X_2 + w_{23}X_3 \quad (4)$$

$$Y = w^T X \quad (5)$$

Basically, regression model training is a process to obtain weight and bias values that minimize the loss function, which usually takes MSE (Mean Square Error) as the loss function in machine learning modeling given in the following formula:

$$MSE = \frac{1}{2n} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 \quad (6)$$

The MSE value was optimized using ordinary least squares and can be obtained as an analytical (close form) solution. However, this analytical formula will be problematic if there is strong multi-collinearity between the predictor variables [38-39]. The MSE of a multi-output system is similar to the MSE in Eq. (5) where each Yi and the associated prediction have at least 2 values.

If there are large predictor variables in the multiple regression model, a penalty will be given to the MSE loss function so that the new model is called ridge regression having the loss function formula as the following.

$$MSE_{ridge} = \frac{1}{2n} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 + \alpha \|w\|_2^2 \quad (7)$$

Eq. (6) is a loss function that must be minimized and is a non-linear function in the w parameter and also contains the alpha hyperparameter [40-41]. In this research, various learning algorithms and alpha hyperparameter values were



tested. The combination of the learning algorithm and the alpha value that produces the minimum loss function of the ridge regression is selected as the model's hyperparameters.

C. Multi-layer Perceptron Neural Network

A neural network is known as a reliable non-linear model for modeling a complex system. The main difference with a multiple regression model is that the weights on the neural network cannot be interpreted as in the multiple regression model, but the magnitude of the weights only indicates the strength or weakness of the relationship between two adjacent nodes. In addition, in the neural network model, each node uses a certain formula called the activation function which is generally a non-linear function [42]. The diagram of an MLP neural network model is illustrated in Fig. 3 as follows:

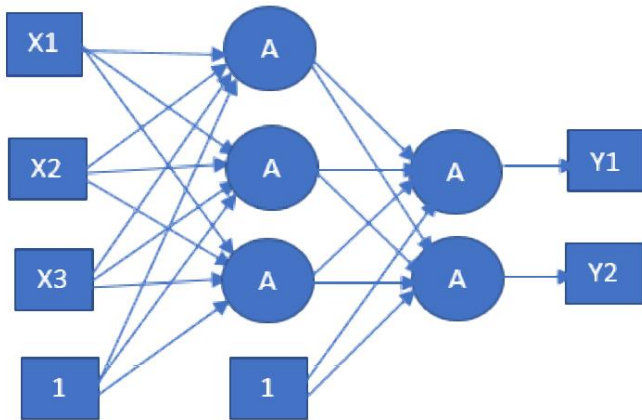


Fig. 3. The Multioutput MLP Neural Network Diagram.

The MLP neural network is characterized by the presence of a hidden layer located between the input layer and the output layer. When this hidden layer is dropped, it will be formed a diagram that is very similar to Fig. 2 except for the existing activation function in the output layer. The neural network model without a hidden layer (very similar to the multiple regression model) is known as the perceptron model. The most popular activation function is known as ReLu which stands for Rectified Linear Units. The ReLu formula is presented in Eq. 7 as follows:

$$ReLU(z) = \begin{cases} z & \text{for } z > 0 \\ 0 & \text{for } z \leq 0 \end{cases} \quad (7)$$

Where  $z$  is a linear combination between the input and the associated weight. For example, the output of the first node on the hidden layer uses the combination linear of  $z = b_1 + w_{11}X_1 + w_{12}X_2 + w_{13}X_3$ , and the first output of the MLP is obtained by using Eq. (8) as the following:

$$Y_1 = ReLu(output\_A1) \quad (8)$$

Where  $output\_A1 = b_h + w_{11} * A_1 + w_{12} * A_2 + w_{13} * A_3$  and  $A_1, A_2, \text{ and } A_3$  are respectively  $ReLU(z1)$ ,  $ReLU(z2)$ , and  $ReLU(z3)$ . In the neural network term, the process calculates the value of the neural network output such as in Eq. (8) is called the forward step.

The loss function of the MLP neural network is the same as the loss function of the multiple regression in Eq. (5) and

Eq. (6) (with and without penalty term). Because the MLP neural network model involves a non-linear activation function such as Relu, the optimal weights for the neural network model cannot be obtained using an analytical solution (close-form solution). The backpropagation algorithm which consists of a forward step and a backward step is used to train this model in obtaining the optimal weights. The forward step aims to calculate the predicted value (network output), while the backward step is the process of updating the weights by applying the gradient descent method and the chain rule to obtain a gradient descent at nodes that are further back [43].

IV. DESCRIBING DATA AND RESEARCH STAGES

This research uses datasets collected by the Center for Child Development Studies at the Midwifery Academy of Wira Husada Nusantara Malang in 2022. In the dataset, there are 696 sample points which are explained by 20 predictor variables and 3 response variables. The predictor variables were derived from several factors including the condition of pregnant women (variables X1 to X11), food intake of pregnant women (variables X12 to X16), and the health condition of the family environment (variables X17 to X20) which the factors were supposed to affect the condition of the baby at the time of birth (variables Y1 to Y3). The predictor variables consist of 12 Likert scales and 8 ratio scales, while the three response variables are all ratio scales. Table I presents the variables in the dataset along with their minimum and maximum values.

TABLE I. THE VARIABLES AND RANGE VALUES

Symbol	Variable	Min.	Max.
X1	Weight at first check	37	80
X2 (Likert)	Frequency of checks during pregnancy	2	4
X3	Weight gain during pregnancy	1	14
X4	Height at first check	139	170
X5	Circumference of the upper arm	18	36
X6	Body Mass Index	16.4	33.3
X7	Normal blood pressure	80	190
X8	Hemoglobin level	5.8	15.4
X9 (Likert)	The protein level in urine	1	4
X10 (Likert)	The number of complaints during pregnancy	1	5
X11	Gestational age when the baby is born	23	42
X12 (Likert)	Consumption of iron element intake	1	4
X13 (Likert)	Consumption of vegetable protein	1	4
X14 (Likert)	Consumption of animal protein	1	4
X15 (Likert)	Consumption of protein from milk intake	1	4
X16 (Likert)	Consumption of vitamin intake	1	3
X17 (Likert)	Family Income per month	1	5
X18 (Likert)	Quantity and quality of drinking water	1	5
X19 (Likert)	Condition of sanitary facilities	1	5
X20 (Likert)	Cleanliness of the house and environment	1	5
Y1	Baby weight at birth	1	4
Y2	Baby body length at birth	30	55
Y3	Baby health score visually	2	10

Differences in the unit of measurement for variables in this dataset must be addressed before modeling is carried out. All variables involved in building the model must have the same unit of measurement (commensurate measures). Therefore, it is necessary to preprocess all ratio-scaled variables using the minimax transformation. By using the formula in equation (1), the value of the transformation results in the range of 0 to 1, so that finally all variables in the dataset have the same unit of measurement.

Furthermore, broadly speaking, the stages of the process to produce the best model in this study are as follows:

- 1) Selecting variables using the filter approach method.
- 2) Splitting the dataset into training and testing subsets.
- 3) Dividing the training subset into k folds and formatting k fold cross validation data.
- 4) Tuning the hyperparameter model using the grid search method.
- 5) Multioutput regression modeling using training subset.
- 6) Ridge regression multi-output modeling with the best hyperparameters using the training subset.
- 7) MLP neural network multioutput modeling with the best hyperparameter using the training subset.
- 8) Evaluating the model performance in both training and testing subsets.

## V. RESULT AND DISCUSSION

The quality of the input data used to build a model significantly determines the model's performance. In this

section, we will discuss variable selection using the filter approach. The multioutput regression model with the least square of the parameter estimation was built based on the training data, the multioutput ridge regression model with the best hyperparameters obtained through cross-validation was built using training data, and the MLP neural networks model with the best hyperparameters obtained through the hyperparameter tuning process was built using the training data. Furthermore, the performance of the models is evaluated on both the training and testing data.

### A. Evaluate Independency among Predictor Variables

In developing a regression model, one of the conditions that must be met is that there is causality between the response and predictor variables. Causality has a meaning that the response variable is influenced by predictor variables. In multiple regression, the predictor variables must also meet the condition that they must be independent of one another. The measuring scale of each variable will determine the appropriate evaluation method to check the independence between the two variables. The independence between the two numerical variables can be evaluated by their correlation value [35]. Table II presents the Spearman correlation value between two predictor variables presented in the form of a matrix. The matrix main diagonal has a value of 1, which indicates the correlation value in the same variable. Because of the limited space, Table II only presents half part of its column elements. The correlation value of two different variables is expressed in the cells outside the main diagonal.

TABLE II. THE SPEARMAN'S CORRELATION BETWEEN 2 PREDICTOR VARIABLES

Variable	X2	X9	X10	X12	X13	X14	X15	X16	X17	X18
X2	1	0.2	-0.13	-0.23	-0.13	-0.26	-0.01	0.03	-0.26	-0.23
X9	0.2	1	0.23	-0.77	-0.68	-0.69	-0.5	-0.63	-0.66	-0.66
X10	-0.13	0.23	1	-0.22	-0.18	-0.19	-0.24	-0.19	-0.22	-0.21
X12	-0.23	-0.77	-0.22	1	0.79	0.75	0.63	0.74	0.8	0.79
X13	-0.13	-0.68	-0.18	0.79	1	0.68	0.58	0.64	0.65	0.66
X14	-0.26	-0.69	-0.19	0.75	0.68	1	0.47	0.55	0.68	0.65
X15	-0.01	-0.5	-0.24	0.63	0.58	0.47	1	0.43	0.56	0.58
X16	0.03	-0.63	-0.19	0.74	0.64	0.55	0.43	1	0.56	0.53
X17	-0.26	-0.66	-0.22	0.8	0.65	0.68	0.56	0.56	1	0.83
X18	-0.23	-0.66	-0.21	0.79	0.66	0.65	0.58	0.53	0.83	1
X19	-0.3	-0.68	-0.21	0.8	0.68	0.72	0.61	0.5	0.81	0.83
X20	-0.26	-0.66	-0.23	0.79	0.68	0.71	0.6	0.5	0.8	0.82
X1	-0.23	-0.34	-0.1	0.43	0.33	0.4	0.26	0.26	0.41	0.41
X3	-0.2	-0.62	-0.23	0.72	0.64	0.68	0.52	0.46	0.75	0.77
X4	0.11	0.23	0.04	-0.28	-0.2	-0.19	-0.19	-0.19	-0.23	-0.22
X5	-0.2	-0.4	-0.15	0.53	0.39	0.47	0.33	0.36	0.49	0.49
X6	-0.25	-0.4	-0.11	0.5	0.38	0.43	0.31	0.32	0.47	0.46
X7	0.12	0.9	0.22	-0.67	-0.6	-0.58	-0.43	-0.58	-0.53	-0.56
X8	-0.2	-0.55	-0.15	0.62	0.51	0.57	0.38	0.45	0.59	0.56
X11	-0.26	-0.49	-0.18	0.56	0.44	0.53	0.42	0.39	0.57	0.57

In this study, two numerical variables are considered independent if the Spearman correlation value is less than 0.7. An evaluation of the correlation values in Table II can be done either by row or column as the basis for selection. Look at the correlation value in column 1 (the variable X2 as the basis), it appears that all correlation values are less than 0.7. It can be interpreted that the variable X2 is independent of all other variables, so X2 is selected as a predictor that has no impact on other variables. Next, the correlation value in column 2 (the variable X9 as the basis) is found 2 correlation values that are greater than 0.7, they are the correlation value between variables X9 and X12, as well as the correlation value between variables X9 and X7. This shows that the three variables, namely X9, X12, and X7 are not independent of each other. The three variables can be represented by one of them as the selected predictor variable. In this research, so that the variable selection process is more structured, the variable that acts as the basis for selection (the variable X9) is determined as the predictor variable. Meanwhile, the row and column associated with variables X12 and X7 are removed or dropped from the correlation matrix member (not considered again in the next predictor variable selection process). The selection process is continued by considering the next column (the variable X10) as the selection bases where there were not find correlation values in the X10's column which are greater than 0.7. The variable X10 is selected as the member of the predictor variables without dropping other rows and columns.

As a summary of the predictor variable selection process in the forwarding next columns given a result that the selection process on the basis of variables X9, X14, X17, and X1 caused as many as 8 variables to be excluded from the set of predictor variables, namely variables X12, X7, X19, X20, X18, X3, X5, and X6. Thus the dataset used to build and evaluate the model in this study consists of 12 predictor variables and 3 response variables. The selected variable rows (12 variables) are the variables that have a role as the basis of the selection, and furthermore they as the predictor variable selected as independent variables or input variables of the model to be built.

### B. Multioutput Regression Model

Initially, the resulting dataset obtained from the selection variables was divided into a training subset (80%) and a testing subset (20%). The training subset data is used to build the model, while the testing subset data is used to evaluate the model's performance. The splitting of the training subset data into five folds aims to form five pairs of sub-training and validation data. The five data pairs will be used for hyperparameter tuning. The emphasis of this research is actually getting a multi-output ridge regression model having a combination of hyperparameters (solver method and alpha value) which produces the smallest MSE value in the

validation data. However, the author also considers it necessary to obtain a multi-output multiple regression model with the ordinary least square estimate as the benchmark model. The multi-output multiple regression model was built based on the training subset data obtained coefficients which are presented in Table III.

On the response variable Y1 (Baby weight at birth), the predictor variable X11 (Gestational age when the baby is born) has a very significant effect (13.893). It is followed by variables X16 (Consumption of vitamin intake), X10 (The number of complaints during pregnancy), and X4 (Weight at first check) which have an effect on the response variable of Baby weight at birth respectively 2.478, 1.907, and 1.662. On the response variable Y2 (Baby body length at birth), the 4 predictor variables with a moderate effect are X9(The protein level in urine), X10 (The number of complaints during pregnancy), X16 (Consumption of vitamin intake), and X15 (Consumption of protein from milk intake) where they have an effect on the response variable of Baby body length at birth respectively 0.16, 0.16, 0.137, and -0.132 respectively. In addition, the 4 predictor variables with a large effect on the response variable Y3 (Baby health score visually) are X14 (Consumption of animal protein), X17 (Family Income per month), X9(The protein level in urine), and X13(Consumption of vegetable protein) with the effect magnitude of -0.993, 0.555, -0.553, and 0.536 respectively. It is clear that the influence of the predictor variables on the response variable of Baby weight at birth is very large, while their influence on the response variable of Baby health score visually is greater than their influence on the response variable of Baby body length at birth.

Before building the multi-output ridge regression model using the training subset data, in this study, hyperparameter tuning (solver method and alpha value) was carried out using 5 folds cross-validation data that had been formed based on the training subset data. For each pair of fold cross-validation data, the model's performance is calculated on the validation data. For example, for solver of 'svd' and alpha of 0.1, parameter estimation is carried out with the 1st sub-training fold data, and then the MSE value is calculated on the 1st fold validation data. The parameter estimation is carried out with the 2nd sub-training fold data and the MSE value is calculated on the 2nd fold validation data. The above computation process is carried out up to the 5th sub-training fold and the 5th fold validation data. So each pair of both solver and alpha was performed five times parameter estimation and five times calculation of MSE value using different sub-training and validation data. Fig. 4 presents the average MSE of each combination of solver and alpha in the validation data that it is presented in the form of a heap map.

TABLE III. THE COEFFICIENTS OF MULTIOUTPUT REGRESSION MODEL

Resp.	X1	X2	X4	X8	X9	X10	X11	X13	X14	X15	X16	X17
Y1	0.231	0.127	1.662	-0.017	0.955	1.907	13.893	0.067	0.893	-0.063	2.478	0.223
Y2	0.004	0.029	0.022	-0.028	0.16	0.16	-0.063	-0.202	-0.09	-0.132	0.137	-0.053
Y3	-0.046	-0.01	0.11	0.219	-0.553	0.004	0.094	0.536	-0.993	-0.025	0.264	0.555

The hyperparameter tuning with grid search method and k-folds cross-validation requires a lot of computation tasks in estimating model parameters on sub-training data and calculating the MSE model performance on validation data. The MSE value in Fig. 4 was obtained from the average of 5 MSE values from five validation data and from 5 models generated from five sub-training data. So in this case, parameter estimation and MSE calculations were carried out 150 times. The smallest MSE average value is 1.561 which occurs at alpha values of 0.1 and 0.3 in all solver methods except the 'sag' solver method which has an MSE value of 1.562. If the MSE value used only considers 2 decimal digits, then all combinations of solver and alpha result in the MSE = 1.56 in all solver methods with an alpha value of less than 0.8.

The multi-output ridge regression model is built by choosing one combination of hyperparameters (solver = 'sag' and alpha = 0.5) having the smallest MSE using the training subset data. The resulted coefficients of the model are presented in Table IV.

The predictor variable having the largest effect on the response variable Y1(Baby weight at birth) is the variable X11(Gestational age when the baby is born). It has a very significant effect of 13.677 which is followed by variables X16(Consumption of vitamin intake), X10(The number of complaints during pregnancy), and X4(Weight at first check). They have an effect on the response variable of Baby weight at birth respectively 2.441, 1.891, and 1.636. For the response variable Y2(Baby body length at birth), the four predictor variables have a moderate effect namely X10(The number of complaints during pregnancy), X9(The protein level in urine), X16(Consumption of vitamin intake), and X15(Consumption of protein from milk intake). They have an effect on the response variable of Baby body length at birth respectively 0.156, 0.155, 0.136, and -0.133. In addition, the four predictor variables with a large effect on the response variable Y3 (Baby health score visually) are X14(Consumption of animal protein), X17(Family Income per month), X9(The protein level in urine), and X13(Consumption of vegetable protein) with the effect magnitude of -0.985, 0.556, -0.554, and 0.544 respectively. It is clear that the influence of the predictor variables on the response variable of Baby weight at birth is very large, while their influence on the response variable of Baby health score visually is greater than their influence on the response variable of Baby body length at birth.

C. MLP Neural Network Model

Neural network modeling is a type of non-linear modeling that is complex because it involves setting two groups of hyperparameters, namely it related to network architecture and it related to network training processes. The hyperparameters in the network architecture include the number of inputs, the number of outputs, the number of hidden layers, the number of

modes in each hidden layer, the activation function employed, the minimized cost function, and others. The hyperparameters in network training include learning algorithms, learning rate values, number of iterations, tolerance values, number of mini-batches, gamma regularization values, and others.

Because the dataset in this study consists of 12 predictor variables and 3 response variables, this leads to neural networks having the architecture of the number of inputs = 12 and the number of outputs = 3. Several hyperparameters were determined by the researcher through a trial and error process, namely the activation function = ReLu, the loss function = MSE, learning algorithm = SGD (stochastic gradient descent with learning rate = 0.01, and momentum value = 0.9), number of iterations = 100, and number of mini-batches = 30. There are two hyperparameters that are considered very important, namely, the number of nodes in the hidden layer and gamma values in L2 norm regularization are determined using the grid search method using the cross-validation data. The variations in the number of hidden nodes that were tested were [12, 18, 30, 42, 60, 78], while the variations in gamma values were [0.001, 0.005, 0.01, 0.05, 0.1, 0.5].

The process of finding the combination of the number of hidden nodes and the gamma value that produces the minimum average MSE value in the validation data is similar to that carried out in the process of obtaining the combination of the solver method and alpha value in multioutput ridge regression modeling. In essence, for each combination of the number of hidden nodes and lambda values, network training is carried out on five sub-training data and the MSE value is calculated for the five corresponding validation data, and finally, the average of the five MSE values obtained is calculated. After the average MSE for all combinations of the number of hidden nodes and gamma value is obtained, then in order to facilitate the process of the grid search method, the average MSE value is presented in a heap map in Fig. 5.

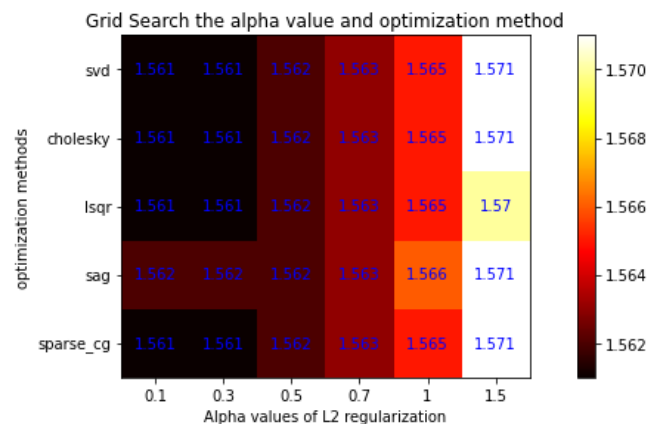


Fig. 4. The Heap Map for Grid Search of the Multi-output Ridge Regression Hyperparameters.

TABLE IV. THE COEFFICIENTS OF MULTIOUTPUT RIDGE REGRESSION MODEL

Resp.	X1	X2	X4	X8	X9	X10	X11	X13	X14	X15	X16	X17
Y1	0.235	0.124	1.541	-0.015	1.002	1.827	12.882	0.093	0.865	-0.048	2.304	0.228
Y2	0.001	0.027	0.02	-0.027	0.138	0.142	-0.071	-0.19	-0.097	-0.135	0.136	-0.05
Y3	-0.046	-0.011	0.113	0.222	-0.554	-0.004	0.12	0.574	-0.951	-0.026	0.269	0.562

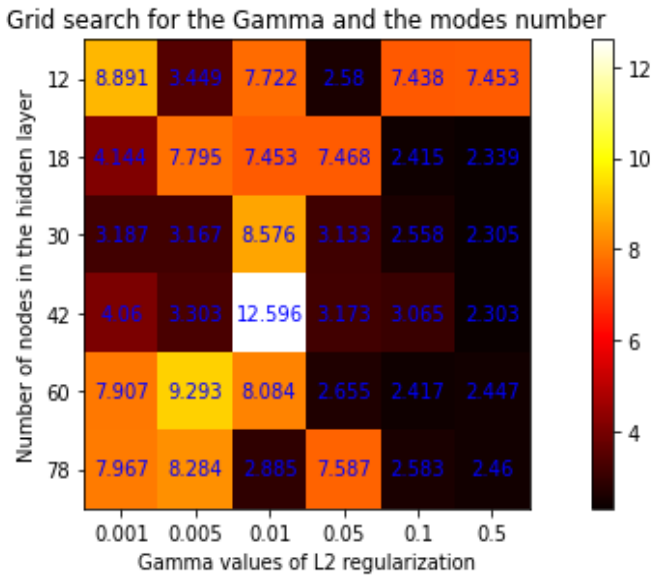


Fig. 5. The Heap Map for Grid Search of MLP Neural Network Hyperparameter.

As explained in the proposed method session, this research uses an MLP neural network architecture whose main feature is that there is only one hidden layer. Based on the average MSE value in the validation data on the heap map in Fig. 5, it is clear that the average MSE value is in the range between 2.415 and 12.596 which is expressed in the darkest (black) and lightest (white) colors. The Heap map also implies that changes in the two hyperparameters greatly affect the MSE average. The combination of the number of nodes and the gamma value that produces the minimum MSE is a combination of the number of nodes = 18 and the value of gamma = 0.1 which will be used to train the network on the training subset data, and then calculate its performance on both the training and testing subset data.

Neural network training with features that include the number of inputs = 12, the number of nodes in the hidden layer = 18, the number of outputs = 3, the number of iterations = 100, and the mini-batch size = 30, the learning algorithm = SGD (learning rate = 0.01 and the momentum value = 0.9) and the value of gamma regularization = 0.1 in the training subset data obtained by the network weight values. The distribution of the resulting network weights in the hidden layer is given in Fig. 6. While the distribution of the resulting network weights in the output layer is given in Fig. 7.

The total number of weights in the hidden layer is  $(12 \times 18) + (1 \times 18) = 234$  where the magnitude of these weights only states the fire strength between each network input and each node in the hidden layer. Based on Fig. 6, it appears that the weight value with the highest frequency (more than 60 pieces) on the average class value = 0.00. The negative weights are about 60 pieces, while the rest are positive weights. Thus, the positive weight dominates in the hidden layer. The effect of the gamma regularization value is to ensure that the weights with very small values become zero, resulting in the number of zeros occupying the mode value of the histogram in Fig. 6.

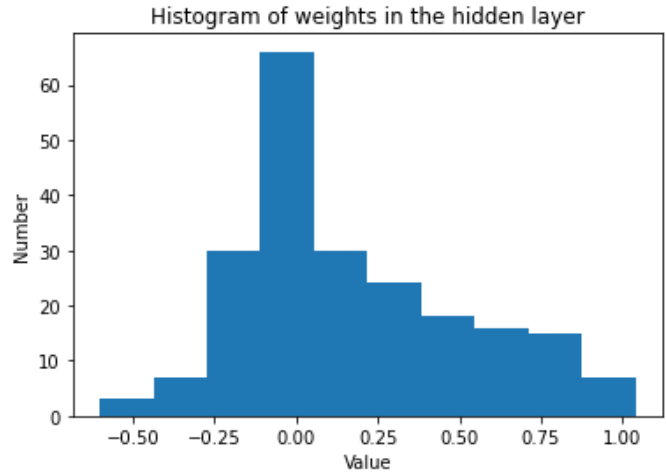


Fig. 6. The Hidden Layer Weights Distribution.

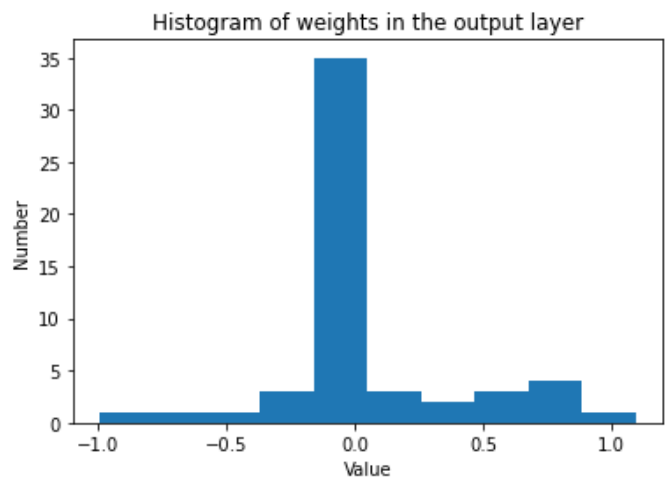


Fig. 7. The Output Layer Weights Distribution.

In the output layer of this neural network model, there are weights as many as  $(18 \times 3) + (1 \times 3) = 57$  pieces. The weights relate the 18 hidden nodes to the 3 output nodes and also relate one bias in the hidden layer to the 3 output nodes. The weights around zero occupy the mode value of the histogram in Fig. 7 which is due to the effect of the l2-norm regularization. The distribution of weights in the output layer is almost similar to the distribution of weights in the hidden layer where positive weights dominate.

#### D. Discussion

In this section, a discussion is given of the results obtained in the previous session, and also the performance of the model is calculated both on the training and testing subset data using the RMSE and R2 measures.

Based on Table V, in general, the coefficient of the ridge regression model has a slightly smaller effect on the response variables than the coefficient of multiple regressions. This is due to the effect of giving the l2-norm regularization value in the ridge regression model. The predictor variable X11 (Gestational age when the baby is born) has a very dominant effect (13.677) on the response variable Y1 (Baby weight at



birth), the predictor variable X9 (The protein level in urine), and X16 (Consumption of vitamin intake) have the greatest influence (0.156) on the response variable Y2 (Baby body length at birth), and the predictor variable X14 (Consumption of animal protein) has the greatest effect (-0.985) on the response variable Y3 (Baby health score visually). The predictor variables X10, X16, and X9 have a considerable influence on two response variables at once. The performance of both regression models and also the MLP neural network model are given in Table VI.

TABLE V. THE COMPARISON OF THE PREDICTOR VARIABLES AFFECTS ON EACH RESPONSE VARIABLE

Response	Predictor	Regr ess	Ridge regress
Y1(Baby weight at birth)	X11(Gestational age when the baby is born)	13.893	13.677
	X16(Consumption of vitamin intake)	2.478	2.441
	X10(The number of complaints during pregnancy)	1.907	1.891
	X4(Weight at first check)	1.662	1.636
Y2(Baby body length at birth)	X9(The protein level in urine)	0.16	0.155
	X10(The number of complaints during pregnancy)	0.16	0.156
	X16(Consumption of vitamin intake)	0.137	0.136
	X15(Consumption of protein from milk intake)	0.132	0.133
Y3(Baby health score visually)	X14(Consumption of animal protein)	-0.993	-0.985
	X17(Family Income per month)	0.555	0.556
	X9(The protein level in urine)	0.553	-0.554
	X13(Consumption of vegetable protein)	0.536	0.544

TABLE VI. THE PERFORMANCE MODEL ON BOTH TRAINING AND TESTING SUBSET DATA

Model	Training		Testing	
	MSE	R2	MSE	R2
Regression	1.5073	0.6469	1.4362	0.6085
Ridge Regression	1.5074	0.6468	1.4336	0.6098
Neural network	2.2714	0.5823	2.1063	0.5574

All of the developed models have similar performance's characteristics which are the RMSE value in the testing subset is smaller than the RMSE value in the training subset. while the R2 value in the training subset is greater than the R2 value in the testing subset. The multiple regression models consistently have better performance than the ridge regression and MLP neural network model in both the training and testing subsets, although the performance difference between the multiple regression and ridge regression models is very small. This result is in contradiction with the level of complexity in the model building where the MLP neural network model

involves as many as 291 weights and also hyperparameter tuning which requires expensive computations. The coefficients of the multiple regression model are obtained based on the close form solution by the ordinary least square method. In another hand, the coefficients of the ridge regression model are obtained using a numerical optimization method that involves several hyperparameters.

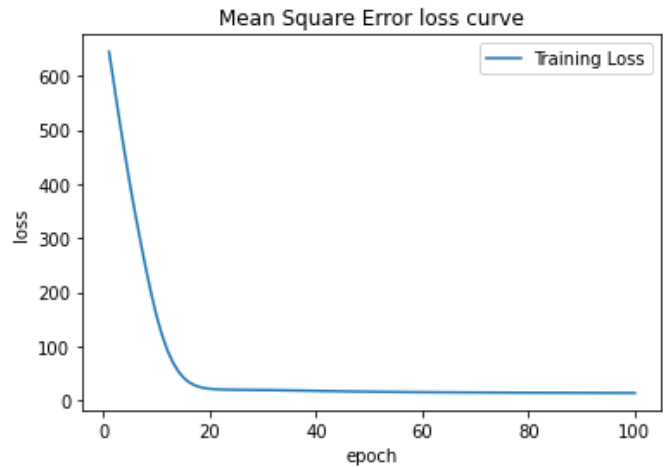


Fig. 8. The Learning Curve of the Logistic Regression Model.

A closed-form solution can be obtained because the predictor variables are independent of each other so the resulting quadratic matrix is not singular. This is one of the impacts of the selection of predictor variables. The outperformance of the multiple regression model also shows that the relationship between the predictor variables and the response variables is based on a linear system. If a linear system is modeled with a non-linear model (such as ridge regression or neural network) it will result in an unsatisfactory performance caused by over modeling. The curve loss function in Fig. 8 strengthens the above argument.

As previously mentioned, the MLP neural network model involves 291 weights that must be optimized using a training subset of 556 instances. The loss function curve in Fig. 8 shows that the value of the loss function has sloped at less than 20 iterations. This means that the model training process is very fast, which only requires updating the weights less than 20 times. This indicates that the system being modeled is a linear system so if it is modeled with a non-linear model, it causes a lot of useless resources or an inefficient modeling process which ultimately results in unsatisfactory model performance.

## VI. CONCLUSION

The equivalence of measure units in the dataset must receive careful attention because arithmetic operations on all mathematical formulas can only work if all operands (variables) involved in the formula must be commensurate. The min-max transformation is often applied to satisfy the commensurate nature of the variable. Multi-collinearity between predictor variables must be overcome so that the influence of predictor variables on response variables is unbiased. The Spearman correlation value can be used as a basis for variable selection with a filter approach if the predictor variables are all numerical scale (interval or ratio).



The complexity of a model does not always result in better performance. In this study, the multiple regression model has the best performance compared with the ridge regression and the MLP neural network model. Even the MLP neural network model has the highest RSME and the lowest R2 value compared to the other two models and its performance gap is moderately large. In this dataset, both the predictor and the response variables are manifest variables that construct the variables of predictor and response latent. So it is an interesting idea if in future research this dataset is modeled with another approach such as the structural equation modeling method using the partial least squares algorithm.

#### REFERENCES

- [1] Hadisuyitno J and Riyadi B D, "Determinant Factors of Stunting Events of Toddler in Batu City Indonesia" *Systematic Reviews in Pharmacy*, vol. 12, no. 1, pp. 231-234, 2021.
- [2] Widodo A and Handoyo S, "The Classification Performance Using Logistic Regression And Support Vector Machine (Svm)" *Journal of Theoretical & Applied Information Technology*, vol. 95, no. 19, pp. 5184-5193, 2017.
- [3] Handoyo S, Chen Y P, Irianto G and Widodo A, "The Varying Threshold Values of Logistic Regression and Linear Discriminant for Classifying Fraudulent Firm" *Mathematics and Statistics*, vol. 9, no. 2, pp. 135 – 143, 2021.
- [4] Nugroho W H, Handoyo S, Akri Y J and Sulistyono A D, "Building Multiclass Classification Model of Logistic Regression and Decision Tree Using the Chi-Square Test for Variable Selection Method" *Hunan Daxue Xuebao/Journal of Hunan University Natural Sciences*, vol. 49, no. 4, pp. 172-181, 2022.
- [5] Taiwo I A, Adeleye A and Uzoma I C, "A possible model for estimating birth length of babies from common parental variables using a sample of families in Lagos, Nigeria" *African Health Sciences*, vol. 21, no. 1, pp. 349-356, 2021.
- [6] Santosa A, Arif E N and Ghoni DA, "Effect of maternal and child factors on stunting: partial least squares structural equation modeling" *Clinical and Experimental Pediatrics*, vol. 65, no. 2, pp. 90-102, 2022..
- [7] Sajjad M, Khan S U, Khan., Haq I U, Ullah A, Lee M Y and Baik S W, "Towards efficient building designing: Heating and cooling load prediction via multi-output model" *Sensors*, vol. 20, no. 22, pp. 6419,2020.
- [8] Turetskyy A, Wessel J, Herrmann C and Thiede S, "Battery production design using multi-output machine learning models" *Energy Storage Materials*, vol. 38, pp. 93-112, 2021.
- [9] Barhmi S, Elfatni O and Belhaj I, "Forecasting of wind speed using multiple linear regression and artificial neural networks" *Energy Systems*, vol. 11, no. 4, pp. 935-946, 2020.
- [10] Handoyo S and Chen Y P, "The Developing of Fuzzy System for Multiple Time Series Forecasting with Generated Rule Bases and Optimized Consequence Part" *International Journal of Engineering Trends and Technology*, vol. 68, no. 12, pp. 118-122, 2020.
- [11] Handoyo S, Chen Y P, Shelvi T M and Kusdarwati H, "Modeling Vector Autoregressive and Autoregressive Distributed Lag of the Beef and Chicken Meat Prices during the Covid-19 Pandemic in Indonesia" *Hunan Daxue Xuebao/Journal of Hunan University Natural Sciences*, vol. 49, no. 3, pp. 220-231, 2022.
- [12] Assaf A G, Tsionas M and Tasiopoulos A, "Diagnosing and correcting the effects of multicollinearity: Bayesian implications of ridge regression" *Tourism Management*, vol. 71, pp. 1-8, 2019.
- [13] Salgado C M, Dam R S F, Salgado W L, Werneck R R A, Pereira C M N A and Schirru R, "The comparison of different multilayer perceptron and General Regression Neural Networks for volume fraction prediction using MCNPX code" *Applied Radiation and Isotopes*, vol. 162, pp. 109170, 2020.
- [14] Tso W W, Burnak B and Pistikopoulos E N, "HY-POP: Hyperparameter optimization of machine learning models through parametric programming" *Computers & Chemical Engineering*, vol. 139, pp. 106902, 2020.
- [15] Belete D M and Huchaiah M D, "Grid search in hyperparameter optimization of machine learning models for prediction of HIV/AIDS test results" *International Journal of Computers and Applications*, pp. 1-12, 2021.
- [16] Marji, Handoyo S, Purwanto I N and Anizar M Y, "The Effect of Attribute Diversity in the Covariance Matrix on the Magnitude of the Radius Parameter in Fuzzy Subtractive Clustering" *Journal of Theoretical and Applied Information Technology*, vol. 96, no.12, pp. 3717-3728, 2018.
- [17] Handoyo S and Kusdarwati H, "Implementation of Fuzzy Inference System for Classification of Dengue Fever on the villages in Malang" *IOP Conf. Series on The 9th Basic Science International Conferences*, vol. 546, no. 5, pp. 052026, 2019.
- [18] Nugroho W H, Handoyo S and Akri Y J, "An Influence of Measurement Scale of Predictor Variable on Logistic Regression Modeling and Learning Vector Quantization Modeling for Object Classification" *Int J Elec & Comp Eng (IJECE)*, vol. 8, no. 1, pp. 333-343, 2018.
- [19] Kusdarwati H and Handoyo S, "Modeling Treshold Liner in Transfer Function to Overcome Non Normality of the Errors" *IOP Conf. Series on The 9th Basic Science International Conferences*, vol. 546, no. 5, pp. 052039, 2019.
- [20] Handoyo S, Marji, Purwanto I N and Jie F, "The Fuzzy Inference System with Rule Bases Generated by using the Fuzzy C-Means to Predict Regional Minimum Wage in Indonesia" *International J. of Oper. and Quant. Management (IJOQM)*, vol. 24, no. 4, pp. 277-292, 2018.
- [21] Utami H N, Candra and Handoyo S, "The Effect of Self Efficacy And Hope on Occupational Health Behavior in East Java of Indonesia" *International Journal of Scientific & Technology Research*, vol. 9, no.2, pp. 3571-3575, 2020.
- [22] Debbouche N, Ouannas A, Batiha I M and Grassi G, "Chaotic dynamics in a novel COVID-19 pandemic model described by commensurate and incommensurate fractional-order derivatives" *Nonlinear Dynamics*, pp. 1-13, 2021.
- [23] Wang D, Zhang Z, Bai R and Mao Y, "A hybrid system with filter approach and multiple population genetic algorithm for feature selection in credit scoring" *Journal of Computational and Applied Mathematics*, vol. 329, pp. 307-321,2018.
- [24] Nayak S K, Rout P K, Jagadev A K and Swarmkar T, "Elitism based multi-objective differential evolution for feature selection: A filter approach with an efficient redundancy measure" *Journal of King Saud University-Computer and Information Sciences*, vol. 32, no. 2, pp. 174-187, 2020.
- [25] Xu D, Shi Y, Tsang I W, Ong Y S, Gong C and Shen X, "Survey on multi-output learning" *IEEE transactions on neural networks and learning systems*, vol. 31, no. 7, pp. 2409-2429, 2019.
- [26] Li X, Luo D, Cheng Y, Wong K Y and Hung K, "Identifying the Primary Odor Perception Descriptors by Multi-Output Linear Regression Models" *Applied Sciences*, vol. 11, no. 8, 3320, 2021.
- [27] Shams S R, Jahani A, Kalantary S, Moeinaddini M and Khorasani N, "The evaluation on artificial neural networks (ANN) and multiple linear regressions (MLR) models for predicting SO2 concentration" *Urban Climate*, vol. 37, pp. 100837, 2021.
- [28] Siavash N K, Ghobadian B, Najafi G, Rohani A, Tavakoli T, Mahmoodi E and Mamat R, "Prediction of power generation and rotor angular speed of a small wind turbine equipped to a controllable duct using artificial neural network and multiple linear regression" *Environmental research*, vol. 196, pp. 110434, 2021.
- [29] Stephenson W, Frangella Z, Udell M and Broderick T, "Can we globally optimize cross-validation loss? Quasiconvexity in ridge regression" *Advances in Neural Information Processing Systems*, vol. 34, pp. 24352-24364, 2021.
- [30] van de Wiel M A, van Nee M M and Rauschenberger A, "Fast cross-validation for multi-penalty high-dimensional ridge regression" *Journal of Computational and Graphical Statistics*, vol. 30, no. 4, pp. 835-847,2021.
- [31] Meanti G, Carratino L, De Vito E and Rosasco L, "Efficient Hyperparameter Tuning for Large Scale Kernel Ridge Regression" In

- International Conference on Artificial Intelligence and Statistics pp. 6554-6572, 2022.
- [32] Blume S, Benedens T and Schramm D, "Hyperparameter Optimization Techniques for Designing Software Sensors Based on Artificial Neural Networks" *Sensors*, vol. 21, no. 24, pp. 8435, 2021.
- [33] Linder A D and Wolfinger R D, "Forecasting with gradient boosted trees: augmentation, tuning, and cross-validation strategies: Winning solution to the M5 Uncertainty competition" *International Journal of Forecasting*, 2022.
- [34] Il Choi H, "Assessment of aggregation frameworks for composite indicators in measuring flood vulnerability to climate change" *Scientific Reports*, vol. 9, no. 1, pp. 1-14, 2019.
- [35] Handoyo S, Pradianti N, Nugroho W H and Akri Y J, "A Heuristic Feature Selection in Logistic Regression Modeling with Newton Raphson and Gradient Descent Algorithm" *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 3, pp. 119-126, 2022.
- [36] Cai J, Luo J, Wang S and Yang S, "Feature selection in machine learning: A new perspective" *Neurocomputing*, vol. 300, pp. 70-79, 2018.
- [37] Elgeldawi E, Sayed A, Galal A R and Zaki A M, "Hyperparameter tuning for machine learning algorithms used for arabic sentiment analysis" *In Informatics*, vol. 8, no. 4, p. 79, 2021.
- [38] Arashi M, Saleh A M E and Kibria B G, "Theory of ridge regression estimation with applications" *John Wiley & Sons*, 2019.
- [39] Liu H, Cai J and Ong Y S, "Remarks on multi-output Gaussian process regression" *Knowledge-Based Systems*, vol. 144, pp. 102-121, 2018.
- [40] Handoyo S and Marji, "The Fuzzy Inference System with Least Square Optimization for Time Series Forecasting" *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*, vol. 7, no. 3, pp. 1015-1026, 2018.
- [41] Marji and Handoyo S, "Performance of Ridge Logistic Regression and Decision Tree in the Binary Classification" *Journal of Theoretical & Applied Information Technology*, vol. 100, no. 13, pp. 4698-4709, 2022.
- [42] Hwang J, Lee J and Lee K S, "A deep learning-based method for grip strength prediction: Comparison of multilayer perceptron and polynomial regression approaches" *Plos one*, vol. 16, no. 2, e0246870, 2021.
- [43] Plonis D, Katkevičius A, Gurskas A, Urbanavičius V, Maskeliūnas R and Damaševičius R, "Prediction of meander delay system parameters for internet-of-things devices using pareto-optimal artificial neural network and multiple linear regression" *IEEE Access*, vol. 8, pp. 39525-39535, 2020.

# Decentralized Access Control using Blockchain Technology for Application in Smart Farming

Normaizeerah Mohd Noor<sup>1</sup>, Noor Afiza Mat Razali<sup>2\*</sup>,  
Nur Atiqah Malizan<sup>3</sup>, Muslihah Wook<sup>5</sup>, Nor Asiakin  
Hasbullah<sup>6</sup>

Defence Science and Technology Faculty  
National Defence University of Malaysia  
Sungai Besi, Kuala Lumpur Malaysia.

Khairul Khalil Ishak<sup>4</sup>

Center of Cyber Security and Big Data  
Management and Science University  
Shah Alam, Selangor, Malaysia

**Abstract**—The application of the Internet of Things (IoT) plays a crucial role in the fourth industrial revolution. The sophistication of technology due to the integration of heterogeneous smart devices open a new threat from various aspects. Access control is the first line of defence to ensure that IoT resources are secure by preventing illegitimate users from gaining access to these resources. However, access control mechanisms face the limitation of technology in large scale IoT deployments since they are based on a centralized architecture. Significant research concerning decentralized access control solutions for securing IoT resources using combined techniques, such as blockchain, have caught much research attention in recent years. Nevertheless, research for decentralized access control for application in smart farming domain remain as a gap. Thus, this study presented a structured literature review on 81 articles related to the field of access control in IoT and blockchain technology to understand the challenges of centralized access control in securing IoT resources. This study serves as a foundation for decentralized access control using blockchain technology and its application to ensure the IoT actuators and sensors security with the aim to be applied in smart farming. This paper was deliberated based on systematic literature review that was searched from four different database platforms between 2018 and 2021. This study mostly addresses the relevant techniques/approaches including blockchain technology, access control model, key management mechanism and the combination of all three methods. The possible impacts, gap, procedures and evaluation of the decentralized access control are highlighted along with major trends and challenges.

**Keywords**—Blockchain; access control; smart contract; internet of thing

## I. INTRODUCTION

Smart farming is the technology enabler that support food security [1] and it has brought changes that reduce costs and minimise environmental constraints, thereby boosting production productivity [2]. Smart farming is capable in enhancing the quality and quantity of production, predicting any possible crop diseases, while optimising agricultural resources and its process. Technological advancement plays a vital role in catalysing the transformation of the smart farming [3]. Data collection using IoT devices such as actuators, sensors, drones and robots are connected to the network for real time data transmission to assist operations. However, the new norm of devices connectivity opened security and privacy

risks for device-based services. Unauthorised access to the devices is among the risk. The situation can be controlled by secure mechanism for authentication in all devices or connected systems. Access control and authentication are considered to be the first lines of defence in restricting unauthorised users from gaining access to IoT resources that provide the data to the smart farming ecosystem. Authentication enables legitimate users to access resources in an authorised manner [4] supporting by access control as the main mechanism for authentication and authorisation, as well as the authority to control resources [5]. Authentication will guarantee that only authorised users are allowed to access a resource. In IoT, access control assigns different privileges to various users regarding the resources of a wide IoT network [6]. However, most existing IoT systems have adopted a conventional access control method which relies on a centralized approach. This may lead to a single point of failure or performance bottlenecks. For instance, an attacker can act as an administrator by stealing authority to illegally access resources, causing a lack of confidence and integrity in such systems. Centralized systems can also be utilised to allow device tracking or related activities, which may compromise privacy. As the number of connected devices increases, it is difficult to manage massive numbers of devices in collecting and handling data using the traditional centralized approach. As a result, IoT has created a challenge in adopting centralized management since it is unable to cope with a large-scale system due to heterogeneous IoT and scalability issues [7], leading to frequent bottlenecks. Researchers have found that applying the blockchain technology can be an alternative solution to this issue. However, the adaptation of decentralized access control in smart farming requires further study to estimate the optimum level of adaptation. Based on the problem statement deliberated above, this research presents a systematic literature review (SLR) on the decentralized access control method using blockchain due to the importance of decentralized access control to manage the heterogeneity and expansion of IoT resources and their application in various domains especially in the domain of smart farming.

The contribution of this review paper is to provide extensive review of research articles to determine the existing gaps, methods and techniques in applying decentralized access control to secure IoT resources. The aim of this study is to

Corresponding Author\*  
FRGS/1/2021/ICT07/UPNM/02/1

understand the current state of related work to address the following research questions:

RQ1: What are the gaps in current access control systems within the IoT ecosystem which can be enhanced by applying blockchain technology?

RQ2: What methods/techniques/approaches are suitable for enhancing access control within the IoT ecosystem by applying blockchain technology?

RQ3: How the evaluation was done to determine the effectiveness of methods/techniques/approaches in previous studies.

## II. RESEARCH METHODOLOGY

The research methodology adopted in this study is the structured literature review (SLR). The SLR was conducted using the following methodology as shown in Fig. 1. Four databases (ScienceDirect, IEEE, Springer and ACM) were employed. Search query was performed to obtain the relevant research articles including journals, book chapters, proceeding papers and books. This paper examines all applicable articles related to decentralized access control for IoT environment application. The method described in [8] was utilised to choose the most important articles related to this research objectives. Articles filtering was done using the six filters that are defined in Fig. 1. A total of 8567 articles were gathered using the following search strings: "Access Control", "Decentral\*", "IoT" or "Internet of Thing" and "Blockchain". Then, after second filter was employed, 5964 articles published between 2018 and 2021 was selected. Next, used source types as filters to reduce the number of articles, thereby producing 2562 results. Then selected papers written in English which yielded 2560 articles. The computer science field was chosen according to the abstract of the paper, resulting in 146 articles. After thoroughly reviewing each paper, a final selection of 81 articles was made upon extensive evaluation based on this study's research questions.

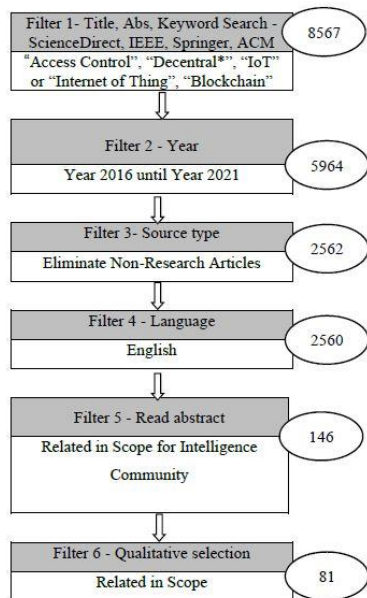


Fig. 1. Filters of the SLR Strategy.

## III. DISCUSSION

Blockchain technology adaptation has been proposed to overcome centralized access control issues. It is expected to be the first line of defence before information sharing of resources is being allowed [9]. Blockchain can also solve security and privacy issues using the decentralized feature by providing security and encryption, making it difficult for attackers since it can detect any illegal changes in its records [10]. To secure resources within an organisation, blockchain technology utilises a cryptography feature that has both a public key and private key that authenticate users who register themselves in the system. A user's personal information is applied to authenticate an individual's identity by employing unique identification, name or biometric data mapped on the user's public key and stored in the blockchain-based smart contracts. Thus, blockchain only provides access to authorised users when accessing resources by authenticating the public key [11].

### A. Gaps of Existing Access Control Systems in the IoT Ecosystem

Fig. 2 presents the research articles regarding access control in IoT by various sectors to determine the gaps to better understand access control issues in the IoT ecosystem. The findings was divided into the several sectors which include smart cities, smart vehicles, smart grid, smart homes, healthcare, banking, property, industry (manufacturing and construction) and general IoT [7], [11], [12], [13], [14], [15], [16], [17], [18], [19], [20], [21], [22], [23], [24], [25], [26], [27], [28], [29], [30], [31], [32], [33], [34], [35], [36], [37], [35], [38], [39], [40], [41], [42], [43]. To answer RQ1 formulated in this study, the analysis was done to understand current access control gaps focusing on the IoT ecosystem that is determined based on the SLR outcomes. The issues were categorised into four groups based on the gaps and problem statement of the literatures: 1) common insufficient IoT access control issues, 2) blockchain combination approaches utilised in the literature, 3) cyberattack issues and 4) lack of security and privacy issues. Based on the literature, several researchers have discussed the common problems in the IoT ecosystem which are: centralized architecture, single point of failure, scalability, heterogeneity, mobility and high energy consumption. The problems are added by the nature of IoT actuators and sensors that are resource-constraints with bandwidth limitations for communication and unable to execute high and memory-intensive computation operations. These problems are major challenges that hinder optimum access control [7]. Researchers in [44], [45, 46], [47] proposed the development of design and standards to secure communication protocols that are capable of interfacing existing systems, collecting data generated by IoT resources and exchanging data to solve trust issues among devices.

Currently, most existing solutions for IoT access control have been developed based on conventional access control architectures, mechanisms, models and policies that mainly rely on single server and third-party entities, leading to high possibility in serious information breach. Failure to ensure the effectiveness of access control may lead to access of information by restricted third party [48]. Thus, conventional access control are inadequate for addressing dynamic and

diverse access control requirements for future IoT ecosystems with new emerging capabilities and application [13] that lead to various security risk including exposure to cyberattack [14]. Common cyberattacks comprise of reuse attacks, DDOS attacks [49] and poisoning attacks. These attacks can cause various drawbacks by exploiting and hijacking the system to retrieve sensitive data. In [50], it was found that attackers can capture, steal or duplicate data to perform illegal activities. In [51], the discussion was done regarding single trusted entities that have become more challenging since these centralized security companies may be biased; permitting illegal or transitive requests while denying legal requests. Attackers can destroy, change or misuse sensitive data and sell it for monetary benefits, leading to data disclosure of user security [31] and lack of data integrity.

Various schemes and cryptographic algorithms were proposed to solve security related issues of IoT by researchers [52], [53], [54]. The proposed methods include a hybrid cryptographic algorithm technique capable of substituting conventional cryptographic algorithms. The same level of security can be simultaneously maintained, leading to additional cost and time for completing encryption and decryption processes [44, 47]. However, these techniques are not feasible since the IoT environment has resource constraints such as high computational power and energy consumption of IoT actuators and sensors. The cryptographic method that involve massive data encryption in IoT actuators and sensors that required higher energy consumption also is not possible to be implemented [55].

To eliminate the gap caused by conventional access control, researchers proposed that the combination of access control and blockchain technology in the IoT ecosystem to resolve issues related to the centralized mechanism. However, IoT network transactions that exceed the capabilities of IoT actuators and sensors, can cause further problems. The complexity of blockchain solutions using the consensus algorithm is beyond the capabilities of IoT actuators and sensors, resulting in constraints in computing and processing and limited bandwidth. Researchers also suggested the consideration of a lightweight key management solution with robust and low resource designs. Based on the deliberations in this section, this study presents a summary of the existing gaps mentioned in previous research in Table I.

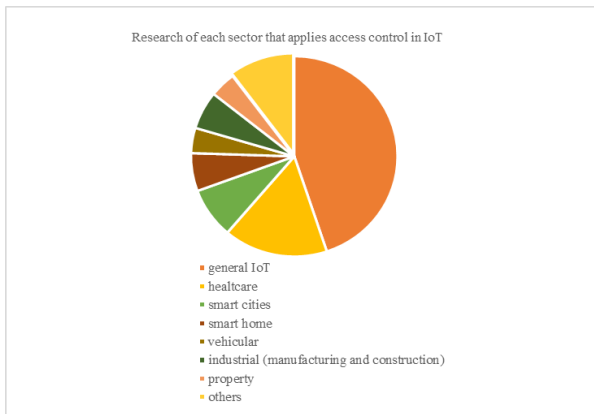


Fig. 2. Access Control Application in IoT for Each Sector.

TABLE I. GAPS IN EXISTING ACCESS CONTROL SOLUTIONS

Current Gaps	Literature Articles
Conventional access control (in centralized architecture) causing single point of failure IoT characteristic related issues (heterogeneity, scalability, mobility, limited power resources, memory size, computational capacity)	[7], [11], [12], [13], [14], [15], [16], [17], [18], [19], [20], [21], [22], [44], [46], [48], [51], [53], [55], [23], [24], [25], [26], [27]
IoT actuators and sensors unable to store large transactions	[23],
Lack of trust and fairness in nodes	[48], [17], [20], [23], [28], [29], [30]
Information leak due to un-restricted access control	[14], [15]
IoT low performance for conventional access control	[31]
Privacy leak risk	[29]
Lack of strong encryption enforcement	[47], [32]
Lack of standardised communication protocol	[47]
Malicious attacks and cyberattacks (including identity spoofing, message eavesdropping, message tampering, content poisoning, physical and cloning attacks)	[11], [45], [46], [14], [49], [50], [20], [25], [33], [34]
Lack of authentication mechanism	[45], [52], [50], [19], [22], [27], [34], [36], [37], [35]
Resource constrained IoT actuators and sensors	[51]
Security weaknesses and vulnerabilities	[46], [49], [55], [33], [38], [39], [56]
Access control founded on blockchain technology related issues in IoT environment: higher cost, increased transaction delays and scalability	[53], [40], [41]
Complexity of consensus algorithm beyond the capabilities of IoT actuators and sensors	[53]
Lack of access control mechanism efficient for the IoT environment	[38]
Centralized client server structure and management schemes less efficient for IoT environment	[25], [32], [33], [42], [43] [42]
High costs in guarding security by combining multiple security technologies	[46]
Traditional fog/cloud computing issues	[57]
Low efficiency in centralized operating environment	[54]
Insufficient conventional storage	[55]
Lack of communication control in data flow	[13]

From the SLR, it can be concluded that to enhance the access control framework in IoT ecosystems, it is crucial to further investigate the following mechanism: access control, trust, elimination of third parties, authentication, privacy and security. The trade-off between the decentralised access control supporting technologies with computing and processing power are vital to be further researched to find the optimum solution.



### B. Techni[58]Ques and Approaches for Existing Decentralised Access Control in IoT

To employ decentralized access control in the IoT ecosystem, researchers have recently applied various techniques to solve the issues presented when addressing RQ1, which are: trust, communication, third party entities, privacy and security problems. Next, evaluated articles were discussed to address RQ2. Based on the literature review, all research papers have presented the use of either blockchain technology, access control models, key management or the combination of all three approaches to efficiently manage access control for IoT resources. The results revealed that only 10 papers used blockchain technology as a strategy for the decentralized access control, while another 10 applied the access control model by adopting the blockchain technology. A total of 12 papers had combined blockchain technology and key management, while three papers combined all three techniques (blockchain technology, access control model and key management techniques). Other techniques were also discussed in these papers, such as IOTA and tangle technology. IOTA is a protocol for securing data communication between IoT actuators and sensors with lightweight quantum resistant cryptocurrency devices. Tangle is an open-source distributed ledger similar to the blockchain technology [21]. Fig. 3 provides a summary of the reviewed papers in this study according to the type of decentralized access control approaches used.

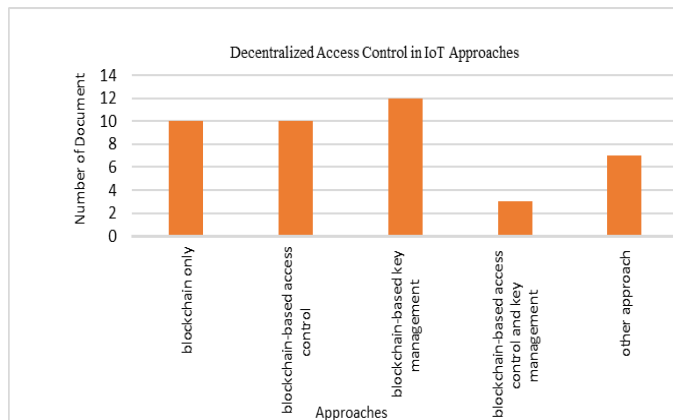


Fig. 3. Type of Decentralized Access Control Approaches in IoT for Each Sector.

1) *Blockchain technology application in decentralized access control in IoT actuators and sensors:* From the literature review, 10 articles were found to have utilised blockchain technology as an approach for decentralized access control in IoT ecosystems. Blockchain is more suitable for decentralized access control due to its immutability and distributed ledgers. It can also handle access control without relying on third parties. Table II presents a summary of blockchain technology employed in decentralized access control including the objectives and techniques used. According to [59], blockchain can be categorised into two groups of access control: 1) global access control and 2) local access control. In global access control, blockchain operates as a distributed ledger and also employs smart contracts to

perform global access control tasks including authentication, authorisation, and key management according to the access control policy. Nguyen et al. proposed a framework for establishing a trustworthy access control mechanism on a mobile cloud platform utilising smart contracts [15]. This study applied blockchain to develop decentralized interplanetary file system (IPFS) on a mobile cloud platform by granting access permissions to each individual medical user to access resources in the environment. The authors further employed IPFS smart contracts to strengthen the security of decentralized cloud storage and data sharing control for better user access management. In the industry domain, Xiong et al. proposed a secure and fair coordinated recognition scheme for multiple IoT actuators and sensors using peer-to-peer edge device cooperation [29]. The study further suggested that using smart contracts can be beneficial in the interaction mechanism of trusted nodes by verifying the node. Among the verification mechanisms, the public key is used to authenticate the digital signature for each node in the environment.

In the second group of local access control, the blockchain is utilised as a distributed ledger that stores access control and verification rules, while the local storage maintains authentication and authorisation. Most researchers focus on the local access control which only stores in blockchain server and uses hash techniques to authenticate and authorise the user and device in the IoT ecosystem. The authors in [50,55] and [35] recommended a lightweight authentication by using lightweight cryptographic key to improve security in IoT actuators and sensors, such as the Merkle-Tree, Streebog Lightweight Hashing Algorithm and Hash-locks. Narayanan et al. proposed the Streebog Lightweight Hashing Algorithm hash generation for faster data encryption. The SALSA20 algorithm was also deemed suitable for the IoT environment in [55] since it can minimise the time consumption. The authors in [50] and [60] used the token mechanism as the access control strategy to authenticate each IoT resource and user. Generally, token consists of unique credentials such as addresses, IDs as well as public and private keys. The research claimed that this approach can reduce computational overhead, time costs for blockchain and enhance efficient access control in the IoT environment. The authors further proposed the combination of public and private blockchain for decentralized authorisation in IoT actuators and sensors. It was claimed that the combination approach can reduce the delay of transaction requests and the amount of data the client requires to send to cloud.

2) *Blockchain based access control:* From the SLR, nine articles were found to adopt an access control model implanted with blockchain technology to provide more fine-grained access control using smart contracts in the IoT ecosystem. Researchers mainly used the distributed attribute-based access control (ABAC) model, capability-based access control (CBAC) model, delegation model, XAML policies and access control list (ACL) model as the proposed approaches for decentralized access control strategy in determining access control of a particular IoT device service through blockchain



and smart contracts. Table III presents the common access control models that were adopted in previous research and their function.

TABLE II. BLOCKCHAIN TECHNOLOGY APPROACH IN DECENTRALIZED ACCESS CONTROL

Authors	Application Objectives	Techniques
[29]	To propose multiple IoT actuators and sensors cooperation driven by secure and fair coordinated recognition scheme using peer-to-peer edge devices	Blockchain, smart contract, YOLO algorithm
[14]	To propose a robust blockchain-based lightweight distributed architecture by leveraging high speed network infrastructure to disburse the computing platform	Blockchain, smart contract, hashing, symmetric encryption, digital signature
[4]	To propose a blockchain based high-efficiency access control framework by leveraging token technology	Blockchain, smart contract, access token, IPFS
[9]	To propose an architecture utilizing blockchain technology in IoT-based environments for healthcare	Blockchain, Hyperledger, chaincode, IPFS
[50]	To propose a blockchain-based authentication mechanism for IoT actuators and sensors	Blockchain, smart contract, Ethereum, token, digital signature algorithm (ECDSA)
[55]	To propose the application of blockchain in enabling secure data sharing among authorised users and devices in the cloud-IoT environment	Blockchain, Streebog lightweight hashing algorithm, SALSA20
[61]	To propose a secure data sharing and access control scheme for users to control the right and privacy of their digital footprint	blockchain, smart contract
[35]	To propose a secure and lightweight Blockchain based IoT authentication scheme	Blockchain, merkle-tree, sequence numbers (SN)
[43]	To propose a novel model for decentralized authorisation by considering limitation of constrained of IoT actuators and sensors	Two blockchain (public and private) Hyperledger, smart contract, Hashed Time-Lock Contracts (HTLCs)
[15]	To propose trustworthy access control mechanism with the application of smart contract on a mobile cloud platform	Blockchain, smart contract, IPFS

TABLE III. ACCESS CONTROL MODEL

Access control model	Functions
ABAC [34]	Attribute-based access control, or ABAC, uses real identities as a set of attributes representing access control policies in a fine-grained method.
RBAC [62]	Role-based Access Control, or RBAC, adopts "roles" as a method to assign permissions. Users are assigned associated role prior to the permission assignment.
ACL	Users of a specific resource will be directly assigned permission in Access control lists (ACLs).

Hossein et al. stored and retrieved data sharing of healthcare records with user-centric and fully distributed architecture of access control, removing trusted third parties in the system [51]. The authors employed different chains of access control policies to ensure that the access policies for the owners of the data are not tampered with, and access to patient data is restricted. The architecture proposes the utilisation of Cluster Head (CH) and Proof-of-Authority (PoA) consensus algorithm to increase the blockchain network throughput and improve system performance and scalability. This approach can reduce the time delay and decrease the number of nodes stored in a single transaction since only miners of each cluster will be kept. However, due to the decreasing number of miners, the risk of malicious activities can increase. In [19], a delegation model approach was employed by adopting blockchain technology for access control in the IoT ecosystem environment setting. The authors proposed an authorisation and delegation model for IoT-cloud based on blockchain technology by deploying smart contracts. The study outcomes indicate that the suggested approach has limitations. The delegation deletion module was not successful since gas requirements exceed the gas limit of the network and further research is required.

Although some limitations do exist, the uniqueness of blockchain technology has attracted technology providers and researchers. One unique feature is the smart contract which can self-execute certain programming conditions and eliminate the need for a trusted entity in the system [63]. Most studies implemented smart contracts to authenticate ownership or to function as a mechanism for controlling token access stored in blockchain. If IoT actuators and sensors are successfully authenticated and validated, devices can access the entire system based on pre-determined access level. Technically, when all authentications are automatically triggered by smart contracts, credibility and impartiality of authentication are theoretically guaranteed. For instance, [64] proposed a trust-based access control framework for decentralized IoT network by applying smart contract to enable decentralization. The authors deployed the ABAC mechanism to manage and limit resources accessed by any party under decided conditions based on the access policy that was set with pre-determined attributes. Access policy enforce smart contracts by assessing the incoming authorisation request to access resources based on context. The context was pre-set with the rulesets according to the specific Boolean attribute in the access policy. Successful authorisation will be followed by a process in generating an access token by smart

contracts. The token can be used to access the resources without repeating the authorisation process for the next access multiple times. This approach provides scalability and at the same time acting as a defender towards Sybil attacks and newcomer attacks that apply attribute registration mechanism way of attack. A similar approach was adopted in [17] by implementing smart contract that functioned as a smart policy to the access control policy. During the execution, a smart policy is created by the resource owner and stored on a blockchain after a proper transaction being transacted.

Meanwhile, in [65], blockchain and decentralized identifier (DID) techniques were used to manage identity and access control for IoT device authentication. This paper deliberated that, based on the proposed mechanism that the capability tokens play a vital component when a particular IoT device service is requesting to obtain the authorisation, the device owner must claim their ownership via the ownership management module to obtain authorisation using the capability token. This approach was determined as lightweight due to three smart contracts applied within the core components: DID registry, device ownership credential registry and device capability credential registry. However, related services need to be present to invoke the functions of these contracts. On the other hand, the work in [66] presented a mechanism where two entities were introduced to handle the delegation process which are labelled as delegator and the delegatee. The entity that executes the role in transferring the access right is called the delegator. The delegator plays a role as the entity that will perform the transfer of the access right, while delegate play a role as the receiving entity. The proposed approach commonly deploys delegation through smart contracts to eliminate the need for a central, trusted, third-party authority. Table IV presents a summary of access control models that adopt blockchain technology in decentralized access control, including the objectives and techniques used are discussed.

3) *Blockchain-based key management for decentralized access control*: From the SLR, 13 articles were found to use the distributed key management in the effort to strengthen IoT access control by applying blockchain to resolve privacy and security issues. Table V presents a summary of blockchain-based key management approaches including the objectives and techniques used for decentralized access control. The combination of distributed key management and blockchain technology is to provide secure authentication and trust communication between device/node in the network layer. The authors in [28] claimed that the use of public key infrastructure (PKI) has vulnerabilities, such as high computational complexity, and requires intermediate certificate authority (CA) to accomplish certificate verification key. Thus, the authors proposed key management in blockchain operated by security access managers (SAMs). SAM plays a crucial role as CA, which is responsible for storing and verifying entire blockchain transactions. Based on SLR, the researchers suggested the adoption of various types of cryptography algorithm techniques including digital signature, endow key trust, symmetric encryption algorithm,

session key, Elliptic Curve Cryptography (ECC), Aggregate signature scheme, Broadcast Encryption (BE) and Multi-Receiver Encryption (MRE) embedded with blockchain. These techniques aim to enhance the access control required for verifying the identity of resources by authenticating and authorising the IoT device and user before entering the system or communicating with other entities in the decentralized nature. For instance, Hammi et al. utilised a bubble of trust in the blockchain environment to provide secure communication to each trusted member device [46]. In this approach, two types of bubbles are present: the master bubble which acts as a certification authority, and the follower bubbles. To authenticate these bubbles, the authors used ECC to generate private/public key-pair since it is known as a lightweight key and is suitable for restricted devices. Smart contract is also applied to verify the uniqueness of the follower's identifier, checking the validity of the follower's ticket using the public key of the master bubble. If one condition is not satisfied, the object cannot be associated to the bubble. If successfully authenticated, the tickets are no longer needed to register new identification and make ACL for users in the system. Shi et al. proposed a blockchain-based access control scheme for privacy preserving in distributed IoT, which formalizes the distributed architecture in IoT and the traditional centralized access control model [25]. The authors utilised domain management server (DMS) to define information and permission of data on blockchain. They used the key-pair of DMS to sign and encrypt data permission on blockchain and employed the symmetric encryption algorithm to encrypt data. Although the data on blockchain is transparent to all nodes, it still reasonably protects the user's privacy.

4) *Blockchain-based access control model and key management*: Based on SLR, 3 articles were found to use distributed key management and access control model that adopted blockchain technology. Various access control models were proposed as access control strategies: RBAC, ABAC, Attribute-Based Signatures (ABS), Anonymous Attribute-Based Encryption (ABE) and Outsourced Attribute-Based Signature (OABS). Double authentication preventing signature (DAPS), Aggregate Signature Scheme, Endow Key Trust, Symmetric Encryption Algorithm and Digital Signature act as the key management for authenticating IoT actuators and sensors. Both techniques were applied together with blockchain technology for enhancing decentralized access control to secure IoT resources as well as improve security and privacy issues in the IoT ecosystem. This combination technique further increased scalability and feasibility of the proposed solution compared to existing solutions.

TABLE IV. BLOCKCHAIN-BASED ACCESS CONTROL APPROACH IN DECENTRALIZED ACCESS CONTROL

Authors	Objective	Detailed Techniques
[22]	To propose a blockchain technology combined with Zero knowledge Token-Based Access Control (BZBAC)	Blockchain, Ethereum, smart contracts, off-chain computation, on-chain, Zero knowledge Token-Based Access Control model
[64]	Trust-based access control framework was developed to support the implementation of decentralized IoT network with smart contracts as the main component	Blockchain, ABAC, smart contract, Trust and Reputation System (TRS)
[51]	To propose a novel access control architecture based on blockchain for storing and retrieving healthcare records	Blockchain, access control policy, cluster head (CH), proof-of-authority (POA) consensus algorithm
[31]	To propose a blockchain-based access control system by embedding ABAC and smart contract on the Hyperledger fabric platform	Blockchain, ABAC policy, smart contracts, Hyperledger fabric
[19]	To propose authentication and delegation mechanisms by using smart contracts and the Stack4Things framework. The mechanism is to support the migration to the decentralized environments	Blockchain, delegation mechanism, RBAC, smart contracts, universally unique identifier (UUID)
[39].	To propose “PrivySharing,” a framework developed based on blockchain aimed to provide secure and privacy-preserving data sharing	Blockchain, smart contracts, ACL rules
[38]	To propose a framework for access control embedded with blockchain technology to enhance privacy policy dedicated for Decentralized Online Social Networks (DOSN)s.	Blockchain, smart contract, ACL rules
[65]	To propose a decentralized capability for IoT access control by implementing blockchain technology with smart contract as the core component	Blockchain, smart contracts, capability based IoT access control, Decentralized Identifier (DiD)
[17]	To present an implementation reference of manipulating XACML policies in a case where Solidity language was used to write a smart contract and deployed on Ethereum platform	Blockchain, smart contracts, XAML policies

Lei et al. proposed a blockchain-based security architecture for improving security and privacy of named data networking (NDN)-based vehicular edge computing (VEC) network. The ABAC mechanism was adjusted into the decentralized architecture; therefore, access control decisions do not have to rely on a centralized policy decision point. The proposed ABAC applies a set of attributes to represent the

resource and the subject requesting the resource [34]. This work also suggested a blockchain-based solution that uses the endow key trust instead of the root key for verifying the authenticity of the key across the user trust domain. Other than that, the symmetric encryption algorithm is also applied to encrypt the content with a symmetric data key and used in access control by controlling the distribution of data key that can only be obtained by an authorised user. Kamboj et al. proposed the RBAC model using blockchain to assign a role in the organisation and management of interactions between users and resources [11]. The role checks and verifies credentials of roles by using smart contracts and digital signature algorithm for signing the transaction and for the generation of public and private keys. Table VI presents a summary of blockchain-based access control and key management approaches, including the objectives and techniques used for decentralized access control.

TABLE V. BLOCKCHAIN-BASED KEY MANAGEMENT APPROACH IN DECENTRALIZED ACCESS CONTROL

Author	Techniques
[40]	Access Control Header (ACH), Cryptographic, multi-layer BC, smart access control
[53]	Blockchain, identity-based signature, hash function, Verifier Control Centre (VCC), Certification Authority (CA)
[25]	Blockchain, symmetric encryption algorithm (SEA), Asymmetric Encryption Scheme (shared key), Management Server (DMS) (Storage)
[44]	blockchain, smart contracts, hybrid cryptosystem with lightweight cryptographic functions (Key Generation Centre (KGC), AES, ECDSA and One-Way Hash Function), angular distance (AD)
[7]	Blockchain, Elliptic Curve Digital Signature Algorithm (ECDSA), Algorithm (Public Key & Private Key), Smart Contract
[48]	blockchain, smart contracts, Elliptic Curve Cryptography (ECC)
[18]	Blockchain, Elliptic curve digital signature algorithm (ECDSA), one-way hash function, session key
[57]	Blockchain, Ethereum, smart contracts, Distributed, Self-Sovereign Identity, fog device authentication mechanism
[33]	Blockchain, cryptographic algorithm- public key, private key and secret key
[42]	Blockchain, Diffie–Hellman, public/private key pair, Session key, Trust Network Framework (TNC), ECDSA
[28]	Blockchain, smart contract, key management schemes, security access managers (SAMs)
[37]	Smart contracts, blockchain broadcast encryption (BE), certificateless multi-receivers encryption (CL-MRE) and Permission Data Hash Table (PDHT)
[46]	blockchain, smart contracts, Public Key Infrastructure (PKI), bubble trust (secure virtual zones), Elliptic Curve Digital Signature Algorithm (ECDSA), ticket

TABLE VI. BLOCKCHAIN-BASED ACCESS CONTROL MODEL AND KEY MANAGEMENT APPROACH IN DECENTRALIZED ACCESS CONTROL

Author	Objective	Detailed Techniques
[11]	Developing a role-based access control method using blockchain technology to manage user-role in the organisation	Blockchain, Ethereum smart contract, RBAC, public key infrastructures (PKIs), elliptic curve digital signature algorithm (ECDSA), digital signature, Keccak-256 cryptographic hash function
[34]	Developing novel security architecture using blockchain technology concept for application in NDN-based VEC networks to address security and privacy challenges	Blockchain, delegate consensus algorithm, access policy key management mechanism (endow key trust, symmetric encryption algorithm), ABAC
[49]	Privacy preserving IoT software update protocol by applying blockchain technology	Blockchain, smart contracts, double authentication preventing signature (DAPS), outsourced attribute-based signature (OABS)

5) *Blockchain with other approaches*: Based on SLR, seven articles were found to employ different approaches to the proposed decentralized access control in the IoT ecosystem. The approaches include Transitive Access Checking and Enforcement (TACE) mechanism which adopts blockchain, blockchain-based access control model, physical unclonable function (PUF), blockchain-based game theory and blockchain-based cross chain technology. Only two articles did not include blockchain adoption. Table VII presents a summary of blockchain combined with other approaches, including the objectives and techniques used for decentralized access control.

TABLE VII. BLOCKCHAIN COMBINED WITH OTHER TECHNIQUES IN DECENTRALIZED ACCESS CONTROL

Author	Detailed Techniques
[47]	Blockchain, smart contracts, Role-Based Access Control, hybrid PUF
[20]	Blockchain, evolutionary combination rule (ECR), smart contracts, game theory
[45]	Blockchain, PUF, smart contracts, Diffie-Hellman key, Chinese Remainder Theorem (CRT), Hash Function
[54]	Blockchain (main chain-consortium), byzantine fault tolerance (RIBFT) algorithm, smart contracts, cross chain technology
[21]	Tangle (store policies), ABAC policy, Decision Point (PDP)
[67]	Blockchain, TACE, Cross-Domain Access Control
[36]	MAM, Tangle, One-Time Signatures (OTS), Merkle Signature Schemes (MSS)

From this review, 39 articles were found to utilise blockchain technology for decentralized access control in IoT ecosystems. Only two articles used IOTA technology similar to blockchain technology. Most research used the ownership concept in the access control model. From SLR extraction, noticed that access control deploys smart contracts to create ownership of resources. The owner will register itself and its resources into smart contracts. After successful registration, smart contracts will generate the credential/token to authenticate the resource owner. The owner can access their resources any time using the credential/token. The deployment of smart contract occurs when two parties agree to the agreement made through coding and can then execute in an autonomous manner. Smart contract is built based on the role or attributes assigned by the authorizing admin who enrolled the smart contract. After deploying smart contract, the user/owner can use their credentials to access the entire network with permission. In a smart contract, several functions are present to operate based on the needs of a contract. Researchers used function add, update, delete and remove to operate in smart contracts. However, the negotiation process for the terms and conditions of smart contracts is unclear.

Access policy is also employed in smart contracts for access control in the IoT ecosystem by creating different levels of user authorisations to access resources. This access policy will be stored in the blockchain server to make it easier for users to invoke their access policy. Authentication and authorisation are also needed in access control for the IoT network. Several techniques that can be used to authenticate and authorise, such as BE, CL-MRE, PDHT, ECDSA and ECC. According SLR, the researchers used public key, private key and secret key to encrypt data for submitting or exchanging data to trusted entities in the IoT network. Several research also used key management to secure communication between device to device (D2D) and device to IoT network. As a result, it is guarded from malicious attacks such as eavesdropping, DDOS and hijacking. From all mentioned techniques, smart contract, authentication & authorisation and key management are the vital components in enhancing the decentralized access control in the IoT ecosystem. However, some techniques are not suitable due to the time delay of transactions and increased overhead. Thus, the trade-off between the techniques and transactions performance must be researched to find the optimum level. Table VIII presents the output of techniques used.

TABLE VIII. OUTPUT OF TECHNIQUES USED IN EXISTING SOLUTION

Author	Techniques	Output
[29]	Blockchain + smart contract	Better fairness and robustness Increased start-up delay
[11]	Ethereum blockchain + new RBAC + PKI + ECDSA	Less execution cost Less running time compared to the RBAC-SC
[47]	Blockchain + access control model + PUF	Cost-effective device in authentication Scalability Computational efficiency of IoT device
[61]	Blockchain + smart contract	Increased feasibility and effectiveness
[4]	Blockchain + smart contract + access token + IPFS	Secure and has low gas cost
[9]	Blockchain + Hyperledger + chaincode + IPFS.	Reduced mining costs and increased throughput
[50]	Blockchain + smart contract ECDSA	More effective in communication overhead compared to previous approach Less time for communication between IoT actuators and sensors with blockchain
[55]	Blockchain, Streebog Lightweight Hashing Algorithm, SALSA20	Better performance Suitable for a large-scale environment Lower time consumption due to spark environment High-level security
[18]	Blockchain + ECDSA + One-way hash function + session key	Low communication cost and access control phases than all existing schemes More computation time than some existing schemes
[20]	Blockchain + game theory	Compared to the environment without the protection shows effectiveness in latency overhead
[33]	Blockchain + public key, private key, secret key	Lower computation cost
[34]	Blockchain + delegate consensus algorithm + key management mechanism + ABAC	NDN: higher throughput in network architecture Time delay: increases total time to verify a transaction signature Increased overhead: encryption and decryption
[42]	Blockchain + public/private key pair + session key + TNC	Longer time to invoke smart contracts Provide stronger mechanism for verification of IoT actuators and sensors that adopt blockchain technology
[43]	Two blockchain (public and private) + Hyperledger + smart contract + HTLC	Decreased overall transaction delay
[15]	Blockchain, smart contract, IPFS	Flexibility in different platforms Availability of data in dynamic real time Decentralized IPFS to solve the single point of failure
[46]	Blockchain + PKI + bubble trust + ECDSA	Less energy and computation consumption
[36]	IOTA + MAM	Less time delay
[37]	Blockchain + BE + CL-MRE + PDHT	Smart contracts increased time cost

#### IV. EXISTING FRAMEWORKS FOR DECENTRALIZED ACCESS CONTROL USING BLOCKCHAIN

Developments in the field of decentralized access control in the IoT ecosystem have attracted various research efforts, resulting in several framework developments based on various objectives and goals. By taking into consideration that authentication and access control are important security aspects, especially with the increase in devices that generate content, various access control solutions have been proposed throughout the literature. In this SLR, found nine existing frameworks for decentralized access control in the IoT ecosystem. The findings are classified into three groups based on framework objectives. Table IX shows the objective regarding existing frameworks.

From the extraction of this SLR, four frameworks were found to develop an access control that focuses on communication control between various entities such as IoT device, gateway, cloud and users [4], [68], [69], [47]. To accomplish the objective of the proposed framework, researchers have adopted blockchain technology to design control communication between various entities by validating data flows before attempting to communicate with other entities. With the capability of blockchain in enhancing reliable communication between entities by utilising its distributed ledger with the hashing function and smart contracts, the authors in [69] designed intra-blockchain interactions within smart contracts. The authors also designed inter-blockchain communication from one node to other nodes and resources in the IoT network. The development of the framework was inspired by the microservice architecture that was build based on 3 proportions: right side, top right side and top left. The core part of this framework is the top right side which utilises 3 smart contracts for IoT systems. The 3 smart contracts have two functionalities: 1) contract level of communication between IoT actuators and sensors, and 2) contract to access data-sources and 3) interoperability of heterogeneous IoT smart contracts. In another approach, the authors in [4] developed access control in various entities and communication control frameworks for cloud-enabled IoT. This framework has three layers: the register model layer, blockchain-based token requesting mechanism layer and requesting data with token used to control the access of users in the system's layers. The authors also deployed pre-defined smart access policies to register resources by the upload mechanism using unique ID. After successful registration, the user must request a token for verifying authority and accessing resources.

In this SLR, two frameworks that focused on authentication and authorisation of user and device, as discussed in [24] and [28] was found. Ma et al. proposed a lightweight, scalable and adaptive key management scheme for the IoT system [28]. In this work, the authors deployed a key management mechanism performed in SAM. The mechanism that was proposed was utilised to record and verify transactions and administrating the key management information. The reason behind the proposed mechanisms is to enable a low-latency key management function for user equipment in the same deployment domain.

Several studies have discussed the constraints regarding devices installed in IoT applications, posing challenges in terms of reliability, cost delay and security. In this SLR, three frameworks that focused on security and privacy enhancements in IoT [70], [71], [72] were found. In the framework proposed in [71], the authors suggested a datagram transport-layer security (DTLS) protocol. The framework was designed with the aim to ensure secure communication that can be realized between three layers: the 1) data producer layer, 2) hybrid computing paradigm layer and 3) data consumer layer. To further strengthen the proposed framework, the authors also included three cryptography mechanisms in the form of algorithms to give higher protection towards system level privacy and security. The proposed combination of blockchain technology and DDSS framework was tested in the decentralized transparent healthcare management system. This framework can be utilised in the application of healthcare domain using a public ledger for each medical record and critical event to provide traceability as well. In addition, in this study, smart contracts usage was applied in automating event-based activities without medical professionals' interference. Meanwhile, in the framework discussed in [73], blockchain technology proposed to be applied in a data-sharing model for intelligent community by utilising the centralized model for access control. The model presented in three modules. In the first module, user authentication and identity management are addressed using enhancement multi-factor authentication model which relies on trusted third parties to manage user authentication. The authors chose not to use blockchain technology in their user authentication module so as to shorten the authentication process and preserve the system's security. However, this approach may lead to various problems in future due to the nature of centralized management. Thus, the gap must be addressed in future work to provide the improvement.

From the literature review analysis, from the observation that the proposed frameworks can be divided into three to five layers based on the physical layer, network layer and application layer concepts. These layers consist of several services and applications in different levels. The first layer is the physical layer, also known as the sensing layer. This layer consists of the IoT device and sensors responsible for collecting and processing data to send to the second layer. Before the IoT actuators and sensors being allowed to enter the network and raw data is transmitted, the access control mechanism will be the first line of defence that guarantee that only eligible actuators and devices will be allowed to access. After the devices clear the access control, then, lightweight key management approach is used to encrypt raw data. The second layer consists of gateway or network paths that are required to transmit IoT data. Any device or user that enters the network must be authorised. Some approaches use simple cryptographic, such as public key, to authorise. Other designs are based on PUF as the key generated for uniquely authenticating IoT actuators and sensors. The third layer is the blockchain layer which performs the transaction validation. This layer uses smart contracts as a core layer that only performs on legitimate devices for accessing resources in the system. Other researchers used a fourth layer as an application

layer which can be executed on cloud or local environments. This layer allows users to access resources by using the internet. To obtain authorisation, users require a valid token or credentials to gain network access.

TABLE IX. BLOCKCHAIN COMBINED WITH OTHER TECHNIQUES IN DECENTRALIZED ACCESS CONTROL

Author	Objectives of the Proposed Framework
[4]	To develop access control in various entities and communication control frameworks for cloud-enabled IoT in terms of data flow from one end to another in CE-IoT services/applications
[47]	To secure data communication and sharing in IoT networks using generated cryptographic keys by providing authenticated device using PUFs and blockchain technology
[71]	To improve the system's security capabilities in classic cloud-centric blockchain-based H-CPS
[70]	To secure and create tamper-resistant massive IoT transactions by improving scalability and the performance of massive IoT networks by utilising blockchain-based secure micro-services in Virtualised Network Functions
[68]	To generate reliable communication IoT eco-systems with reliable information integration between users by validating nodes based on inter-operable structures
[72]	To improve the transaction delay among IoT applications by using blockchain based in SDN architecture
[15]	To allow authorised entities (such as healthcare providers) to effectively retrieve EHRs on cloud, while preventing unauthorised access to EHRs resources
[24]	To enable secure and transparent collaborations for connected IoT actuators and sensors trust-based automation to recognise, authenticate and access control of devices in the perception layer
[28]	To achieve a lightweight, scalable, adaptive key management scheme and authorisation assignment mode by verifying the access query transaction based on logical topology in the IoT system
[69]	To enhance access control traditional development model with features that primarily support intra-blockchain interactions within smart contracts as well as enable inter-blockchain communication to other nodes and resources in the IoT network

## V. EVALUATION FOR DECENTRALIZED ACCESS CONTROL USING BLOCKCHAIN

How the evaluation was done to determine the effectiveness of methods/techniques/approaches in previous studies, is addressed in RQ3. Each reviewed study had been evaluated based on their proposed techniques, approaches and frameworks as the baseline for future investigations. This evaluation was accomplished during the experimental phase. Validation was conducted using pre-determined parameters and by comparing existing baseline models. These parameters and models have been used by numerous research works that reported satisfactory results and were then later examined and enhanced by others. To answer RQ3, this section lists the



datasets, parameters and tools (hardware/software) used to evaluate the performance of the proposed approach.

#### A. Dataset for Evaluation

For every proposal deliberated in the literature, experiments were done to validate the proposals. For the validation, most datasets used in access control experiments generated by nodes. Most nodes are used in the research to simulate the experiment scenario [7,15,31,44,51]. Data were generated by IoT actuators and sensors, such as raspberry pi system [7] sensors, collected from laptops and mobile phones to form a dataset. In [71] and [74], available datasets or open data are employed to conduct experiments. For instance, Guruprakash & Koppu used Kaggle which contains temperature readings from IoT actuators and sensors installed inside and outside anonymous buildings [74]. This dataset was analysed to validate the proposed system functionalities and capabilities [74].

#### B. Parameters for Performance Evaluation

Based on the literature, experiments have been accomplished to evaluate the different performances of the proposed approaches according to various parameters. The parameters frequently depend on the objective of the study and the goal of the experiments. The evaluation in blockchain technology can be categorised into two groups based on the evaluation goals, as follows:

1) *Parameter based on performance of blockchain technology*: The evaluation of blockchain performance metrics and parameters consist of transaction throughput, transaction latency, network latency, block size, computational cost, block validation, storage overhead, transaction delay and time delay [7], [9], [51], [15,25,54], [74], [75], [76]. Network latency is the total time taken for a transaction to be executed in the blockchain network. To evaluate these parameters, Table X displays the measuring units based on the parameters used: milliseconds (ms), second (s), minutes (m), joules (j), ethers, bytes, transaction (Tx), transaction per second (TPS), transaction per minutes (TPM). Based on the study of [15], the time taken is usually higher when the mechanism involved with user authentication is based on smart contracts that consume more time to process user requests, as compared to the non-authenticated scheme. The computation cost based on the time of deploying and invoking a smart contract increases [52]. The storage cost is normally based on the size of the stored data [74]. Transaction throughput is defined as the number of validated transactions per second. According to Zaabar et al., the throughput is separated into two sub-categories: the read throughput and the transaction throughput [9]. The read throughput is defined as the total number of reading operations performed across the blockchain network within the given timeslot, while the transaction throughput is the number of successful transactions performed in the blockchain network within the given timeslot.

2) *Parameter based on performance of access control in blockchain technology*: The evaluation of the performance access control in blockchain were proposed by allowing

authorised entities to effectively retrieve the database and prevent unauthorised access from resources. To verify and authenticate the authorised transactions and un-authenticate unauthorised transactions, several existing solutions consisting of many operations must be accomplished. The authors in [44] highlighted that to execute these operations, the system may consume more energy. For the evaluation of the authentication process, researchers utilised parameters such as energy consumption [44], time taken for encryption and time taken for decryption [52]. To evaluate energy consumption, researchers chose parameters such as cost, time (ms) and energy (j). Regarding the time taken for encryption and decryption, [55] defined the encryption time as the amount of time consumed to convert plaintext into ciphertext, which generally depends on data size and the key size used for encryption. The decryption time was defined as the amount of time taken by the algorithm to convert ciphertext into original data. Storage and communication costs are also parameters in access control. The measuring unit of both parameters is bytes [44].

TABLE X. PARAMETERS AND VARIABLES

Parameter	Variables (unit)
<b>Access Control</b>	
Energy consumption [44]	Cost, time (ms), energy (j)
Time taken for encryption and decryption [55]	Time(ms), data size (bytes)
Storage cost [44]	Cost, size of data key (bytes)
Communication cost [44]	Cost, size of data key (bytes)
<b>Blockchain</b>	
Transaction throughput [77]	Response time (m) and TPM (size of transaction)
Transaction latency	Time (ms) and invoking a transaction (Tx)
Network latency [78],[15]	Time (ms),
Block validation [74]	Processing time (s), number of blocks
Computational cost[79]	Time (s), cost, ethers
Storage overhead [44],[80]	Size of key (bytes), time(s)

3) *Tools for evaluation*: This section provides an overview of the technologies and tools adapted by the articles reviewed in this study. Researchers implemented their proposed solutions by setting up the experimental environment to serve as the underlying functions as well as to efficiently evaluate the proposed solutions or mechanisms and frameworks. The details of the setups are divided into two categories, as follows:

a) *Hardware*: From the extensive review of the selected literature, the commonly used hardware for conducting experiments included desktop pc, laptop, mobile phone, raspberry pi, memory and hard disk. The researchers mainly used large storage and equipment that are compatible with their experiments. The desktop pc and laptops were commonly

employed as the simulation platform and blockchain server to run the experiments. Memory ranging from 8 to 16 GB RAM [44] are necessary [81][34]. Raspberry pi can be used as lightweight IoT actuators and sensors, further acting as IoT nodes. The interaction between IoT nodes were developed using C++ language and the JsonRPC library for communication [7].

*b) Software:* Most studies, as in [59] and [75], chose a private blockchain (such as Ethereum) to develop their blockchain network and conduct experiments. Based on [50], Ethereum is the commonly used platform for building decentralized apps (dApps). It provides a secure way to perform transactions using the elliptic curves cryptography protocol. Ganache is also used to test the decentralized application without an actual set-up of the Ethereum network. Ganache is defined as a blockchain emulator, also known as a personal Ethereum client or node [7]. Several studies have deployed a blockchain network built on Hyperledger fabric to execute experiments, such as in [9,54]. Node.js is also used as an Ethereum network [81]. Many studies further developed an experiment in the virtual environment to build a blockchain network that can be deployed in Ethereum Virtual Machine (EVM), such as in [64]. Some researchers applied a simulator or emulator environment to conduct their experiments. A simulator, such as OMNeT++ [28], can create an environment similar to the original which can configure real devices. An emulator, such as Common Open Research Emulator (CORE) [72], can be used to duplicate all hardware and software features in real devices.

In terms of the programming language, most studies used the python language to create a prototype interface since it is considered to be a dynamic and scalable language across multiple platforms [50]. Web3.py library is frequently employed to enable users to interact with Ethereum clients and request functions written in smart contracts. Solidity programming language is also applied to write smart contracts [7], [10], [27], [62]. These smart contracts were implemented for testing, debugging and then deployment, either in Ethereum Virtual Machine (EVM) [11], Truffle [7], Testnet [62] or Remix IDE [56], before implementing them in the blockchain platform. Ropsten [11,55], Rinkeby and Kovan are Ethereum tools for testing and development purposes. Researchers have noted that benchmarking is important to measure the performance of the blockchain application [9]. The most commonly used benchmarking for the Hyperledger network is Hyperledger Caliper. The use of several appropriate protocols play a crucial role in an experiment. Common communication protocols used are IPV6 and 6LoWPAN [53].

## VI. CONCLUSIONS AND FUTURE WORKS

IoT actuators and sensors are capable to further improve the efficiency of smart farming. However, the security of the IoT actuators and sensors depending on the access control that act as the first line of defence via authentication and authorization. This paper presented the background of decentralized access control in this study. Based on extensive literature review, most commonly applied techniques to

authenticate and authorise users or devices in IoT networks are summarized as key management schemes including the asymmetric cryptographic algorithm, the symmetric cryptographic algorithm, session key, secret key, PKI (including hashing algorithms), Symmetric Encryption, Digital Signature, Elliptic Curve Cryptography (ECC) and Elliptic Curve Digital Signature Algorithm (ECDSA) based on blockchain technology. This approach is vital for securing access control and communication between D2D, user to device and device to network.

Meanwhile, the access control models - RBAC and ABAC, are frequently used to assign a user to a role in the system according to their attribute, credentials or authority to access resources. This approach can be commonly utilised as a strategy for designing various smart contracts for fine grained access control. In the smart contract operation, all information associated with a particular role or attribute will be stored on blockchain. This makes it more transparent and available for other users to access resources with the owner's permission. By deploying the access control strategy in blockchain in the form of smart contracts, the computation overhead of IoT actuators and sensors will be reduced, therefore, the framework can apply lightweight IoT actuators and sensors existed ecosystem.

Based on reviewed articles in this SLR study; tokens were incorporated into the strategy for subjects to obtain access rights by applying the token which can improve access efficiency. Other techniques are also used to make the system more scalable. Researchers commonly use off chain and on chain with other storages called IPFS. The assessment of all proposed techniques was accomplished by establishing the necessary steps to setup the evaluation. The literatures also reported that most research have developed an experimental environment on the Ethereum platform. Some experiments were executed in the virtual environment due to the requirements and needs for large storage and high CPU or laptop processors. The CPU or laptops are used as the main components in an experiment to simulate the blockchain server, or act as a gateway to collect data from IoT actuators and sensors. The core of the development system is smart contracts, which are developed using Solidity programming language. Regarding experiments, datasets were collected using IoT actuators and sensors according to pre-defined parameters for specific experiment designs. Evaluation was then performed to examine the proposed system based on established parameters. The parameters were also used as a baseline comparison with other relevant works and for validating the proposed system.

Among the gaps identified in the current access control data in the IoT ecosystem are: lack of mechanism and standardised protocol of access control and communication protocol, decentralized access control, authentication, privacy and security. From the finding that the mechanism in authorisation and authentication is not fully adapted in a decentralized manner. It remains in the same phase and requires a trusted entity in the validation process. Based on this study, most of the proposed solutions which influence decentralized access control in the IoT ecosystem include a lightweight distributed key management solution, a robust

design in smart contracts, efficient consensus approach and decentralized access control.

This study concludes that the decentralized access control is a relevant topic for researchers to explore and investigate. The combination of access control approaches that adopt blockchain technology can be a possible mechanism for enhancing decentralized access control in the IoT ecosystem. In addition, the access policy based on ABAC and RBAC model can be used to achieve flexibility and dynamic access control using smart contracts. Smart contracts can be used as an automation decision and authorization to eliminate centralized server into decentralized server. The use of multiple layers also plays a crucial role in reducing the scalability of IoT systems, speeding up the process of requesting transactions and reducing time delays. Thus, it is suitable for application in large scale IoT systems that manage big data processing.

Smart farming also relies on the IoT technology and smart systems to collect real-time data and provide observations in management operations on the farm, including pre- and post-harvest. For optimum access control decentralization in smart farming, Ethereum platform that include public and private blockchains can be utilised.

For future studies, in regard to the application of decentralized access control in smart farming, researchers should explore and investigate the enhancement of smart contracts design for access control since smart contracts play a vital role in blockchain. They were designed with the aim to perform event-based automation activities without human interference based on pre-defined contracts. Nevertheless, smart contracts can be the loophole for blockchain technology, which is another gap that must be addressed to further enhance decentralized access control in IoT, especially for the application in smart farming. Thus, the design and mechanism for applying the smart contracts concept in blockchain technology must be further examined to achieve an optimum design. This can be validated through simulations until the establishment of contracts is complete. This is crucial to further secure and strengthen a resource from unwanted threats, including smart contract-related scams and illegal activities.

#### ACKNOWLEDGMENT

The authors would like to acknowledge National Defence University of Malaysia (UPNM) and Ministry of Higher Education Malaysia (MOHE) for the approved fund which makes this research viable and effective. This research is supported by Fundamental Research Grant FRGS/1/2021/ICT07/UPNM/02/1.

#### REFERENCES

- [1] Mat Lazim R, Mat Nawi N, Masroon M H, Abdullah N and Che Mohammad Iskandar M 2020 Adoption of IR4.0 into Agricultural Sector in Malaysia: Potential and Challenges *Adv. Agric. Food Res. J.* 1 1–14.
- [2] Triantafyllou A, Tsouros D C, Sarigiannidis P and Bibi S 2019 An architecture model for smart farming *Proc. - 15th Annu. Int. Conf. Distrib. Comput. Sens. Syst. DCOSS 2019* 385–92.
- [3] Marinchenko T 2020 Digitalization Of Agricultural Sector: Outlook In Russia.
- [4] Chai B, Yan B, Yu J and Wang G 2021 BHE-AC: a blockchain-based high-efficiency access control framework for Internet of Things *Pers. Ubiquitous Comput.*
- [5] Matrazali N, Noor N, Hasbullah N, Chen L, Ishak K and Nordin N 2021 A Conceptual Model: Securing Resources Through a Decentralized Access Control Using Blockchain Technology for Smart Farming pp 399–410.
- [6] Hou J, Qu L and Shi W 2019 A survey on internet of things security from data perspectives *Comput. Networks* 148 295–306.
- [7] Khalid U, Asim M, Baker T, Hung P C K, Tariq M A and Rafferty L 2020 A decentralized lightweight blockchain-based authentication mechanism for IoT systems *Cluster Comput.* 23 2067–87.
- [8] Razali N A M, Malizan N A, Hasbullah N A, Wook M, Zainuddin N M, Ishak K K, Ramli S and Sukardi S 2021 Opinion mining for national security: techniques, domain applications, challenges and research opportunities *J. Big Data* 8 150.
- [9] Zaabar B, Cheikhrouhou O, Jamil F, Ammi M and Abid M 2021 HealthBlock : A secure blockchain-based healthcare data management system *Comput. Networks* 200 108500.
- [10] Zhang Y, He D and Choo K K R 2018 BaDS: Blockchain-based architecture for data sharing with ABS and CP-ABE in IoT *Wirel. Commun. Mob. Comput.* 2018.
- [11] Kamboj P, Khare S and Pal S 2021 User authentication using Blockchain based smart contract in role-based access control *Peer-to-Peer Netw. Appl.* 14 2961–76.
- [12] Rabejaja T, Pal S and Hitchens M 2019 Design and implementation of a secure and flexible access-right delegation for resource constrained environments *Futur. Gener. Comput. Syst.* 99 593–608.
- [13] Bhatt S and Sandhu R 2020 ABAC-CC: Attribute-based access control and communication control for internet of things *Proc. ACM Symp. Access Control Model. Technol. SACMAT* 203–12.
- [14] Deebak B D and AL-Turjman F 2022 A robust and distributed architecture for 5G-enabled networks in the smart blockchain era *Comput. Commun.* 181 293–308.
- [15] Nguyen D C, Pathirana P N, Ding M and Seneviratne A 2019 Blockchain for Secure EHRs Sharing of Mobile Cloud Based E-Health Systems *IEEE Access* 7 66792–806.
- [16] Liu Y, Lu Q, Chen S, Qu Q, O'Connor H, Raymond Choo K K and Zhang H 2020 Capability-based IoT access control using blockchain *Digit. Commun. Networks* 7 463–9.
- [17] Di Francesco Maesa D, Mori P and Ricci L 2019 A blockchain based approach for the definition of auditable Access Control systems *Comput. Secur.* 84 93–119.
- [18] Bera B, Chattaraj D and Kumar A 2020 Designing secure blockchain-based access control scheme in IoT-enabled Internet of Drones deployment *☆ Comput. Commun.* 153 229–49.
- [19] Tapas N, Longo F, Merlino G and Puliafito A 2020 Experimenting with smart contracts for access control and delegation in IoT *Futur. Gener. Comput. Syst.* 111 324–38.
- [20] Esposito C, Tamburis O, Su X and Choi C 2020 Robust Decentralised Trust Management for the Internet of Things by Using Game Theory *Inf. Process. Manag.* 57 102308.
- [21] Shafeeq S, Alam M and Khan A 2019 Privacy aware decentralized access control system *Futur. Gener. Comput. Syst.* 101 420–33.
- [22] Song L, Ju X, Zhu Z and Li M 2021 An access control model for the Internet of Things based on zero-knowledge token and blockchain *Eurasip J. Wirel. Commun. Netw.* 2021.
- [23] Pal S, Rabejaja T, Hitchens M, Varadharajan V, Member S and Hill A 2019 On the Design of a Flexible Delegation Model for the Internet of Things Using Blockchain *IEEE Trans. Ind. Informatics* PP 1.
- [24] Tang B, Kang H, Fan J, Li Q and Sandhu R 2019 IoT passport: A blockchain-based trust framework for collaborative internet-of-things *Proc. ACM Symp. Access Control Model. Technol. SACMAT* 83–92.
- [25] Shi N, Tan L, Yang C, He C, Xu J, Lu Y and Xu H 2021 BacS: A blockchain-based access control scheme in distributed internet of things *Peer-to-Peer Netw. Appl.* 14 2585–99.

- [26] Elahi M M, Rahman M M and Islam M M 2022 An efficient authentication scheme for secured service provisioning in edge-enabled vehicular cloud networks towards sustainable smart cities *Sustain. Cities Soc.* 76 103384.
- [27] Alshahrani M and Traore I 2019 Secure mutual authentication and automated access control for IoT smart home using cumulative Keyed-hash chain *J. Inf. Secur. Appl.* 45 156–75.
- [28] Ma M, Shi G and Li F 2019 Privacy-Oriented Blockchain-Based Distributed Key Management Architecture for Hierarchical Access Control in the IoT Scenario *IEEE Access* 7 34045–59.
- [29] Xiong F, Xu C, Ren W, Zheng R, Gong P and Ren Y 2022 A blockchain-based edge collaborative detection scheme for construction internet of things *Autom. Constr.* 134 104066.
- [30] Ali I, ul Hussen Khan R J, Noshad Z, Javaid A, Zahid M and Javaid N 2020 Secure Service Provisioning Scheme for Lightweight Clients with Incentive Mechanism Based on Blockchain vol 96 (Springer International Publishing).
- [31] Liu H, Han D and Li D 2020 Fabric-iot: A Blockchain-Based Access Control System in IoT *IEEE Access* 8 18207–18.
- [32] Alam M, Emmanuel N, Khan T, Khan A and Javaid N 2018 Secure policy execution using reusable garbled circuit in the cloud *Futur. Gener. Comput. Syst.* 87 488–501.
- [33] Bonnah E and Shiguang J 2020 DecChain: A decentralized security approach in Edge Computing based *Futur. Gener. Comput. Syst.* 113 363–79.
- [34] Lei K, Fang J, Zhang Q, Lou J, Du M, Huang J, Wang J and Xu K 2020 Blockchain-Based Cache Poisoning Security Protection and Privacy-Aware Access Control in NDN Vehicular Edge Computing Networks *J. Grid Comput.* 18 593–613.
- [35] Hong S 2020 P2P networking based internet of things (IoT) sensor node authentication by Blockchain Peer-to-Peer *Netw. Appl.* 13 579–89.
- [36] Brogan J, Baskaran I and Ramachandran N 2018 Authenticating Health Activity Data Using Distributed Ledger Technologies *Comput. Struct. Biotechnol. J.* 16 257–66.
- [37] Lin C, He D, Huang X, Choo K K R and Vasilakos A V. 2018 BSEfn: A blockchain-based secure mutual authentication with fine-grained access control system for industry 4.0 *J. Netw. Comput. Appl.* 116 42–52.
- [38] Ur Rahman M, Guidi B and Baiardi F 2020 Blockchain-based access control management for Decentralized Online Social Networks *J. Parallel Distrib. Comput.* 144 41–54.
- [39] Makhdoom I, Zhou I, Abolhasan M, Lipman J and Ni W 2020 PrivySharing: A blockchain-based framework for privacy-preserving and secure data sharing in smart cities *Comput. Secur.* 88 101653.
- [40] Paul R, Ghosh N, Sau S, Chakrabarti A and Mohapatra P 2021 Blockchain based secure smart city architecture using low resource IoTs *Comput. Networks* 196.
- [41] Alcaraz C, Rubio J E and Lopez J 2020 Blockchain-assisted access for federated Smart Grid domains: Coupling and features *J. Parallel Distrib. Comput.* 144 124–35.
- [42] Zhang J, Wang Z, Shang L, Lu D and Ma J 2020 BTNC: A blockchain based trusted network connection protocol in IoT *J. Parallel Distrib. Comput.* 143 1–16.
- [43] Siris V A, Dimopoulos D, Fotiou N, Voulgaris S and Polyzos G C 2020 Decentralized authorization in constrained IoT environments exploiting interledger mechanisms *☆ Comput. Commun.* 152 243–51.
- [44] Vishwakarma L and Das D 2021 SCAB - IoTA: Secure communication and authentication for IoT applications using blockchain *J. Parallel Distrib. Comput.* 154 94–105.
- [45] Suresh A, Hamza R, Hassan A, Jiang N and Yan H 2020 Computers & Security Efficient privacy-preserving authentication protocol using PUFs with blockchain smart contracts *Comput. Secur.* 97 101958.
- [46] Hammi M T, Hammi B, Bellot P and Serhrouchni A 2018 Bubbles of Trust: A decentralized blockchain-based authentication system for IoT *Comput. Secur.* 78 126–42.
- [47] Satamraju K P and Malarkodi B 2021 A decentralized framework for device authentication and data security in the next generation internet of medical things *Commun.* 180 146–60.
- [48] Huang J C, Shu M H, Hsu B M and Hu C M 2020 Service architecture of IoT terminal connection based on blockchain identity authentication system *Comput. Commun.* 160 411–22.
- [49] Zhao Y, Liu Y, Tian A, Yu Y and Du X 2019 Blockchain based privacy-preserving software updates with proof-of-delivery for Internet of Things *J. Parallel Distrib. Comput.* 132 141–9.
- [50] Hameed K, Garg S, Amin M B and Kang B 2021 A formally verified blockchain-based decentralised authentication scheme for the internet of things vol 77 (Springer US).
- [51] Mohammad Hossein K, Esmaeili M E, Dargahi T, Khonsari A and Conti M 2021 BCHealth: A Novel Blockchain-based Privacy-Preserving Architecture for IoT Healthcare Applications *Comput. Commun.* 180 31–47.
- [52] Zhang Y, Deng R H, Han G and Zheng D 2018 Secure smart health with privacy-aware aggregate authentication and access control in Internet of Things *J. Netw. Comput. Appl.* 123 89–100.
- [53] Fotuhi R and Shams Alikee F 2021 Securing communication between things using blockchain technology based on authentication and SHA-256 to improving scalability in large-scale IoT *Comput. Networks* 197 108331.
- [54] Guo S, Wang F, Zhang N, Qi F and Qiu X 2020 Master-slave chain based trusted cross-domain authentication mechanism in IoT *J. Netw. Comput. Appl.* 172 102812.
- [55] Narayanan U, Paul V and Joseph S 2021 Decentralized blockchain based authentication for secure data sharing in Cloud-IoT: DeBlock-Sec *J. Ambient Intell. Humaniz. Comput.*
- [56] Panda S S, Jena D, Mohanta B K, Ramasubbareddy S, Daneshmand M and Gandomi A H 2021 Authentication and Key Management in Distributed IoT Using Blockchain Technology *IEEE Internet Things J.* 8 12947–54.
- [57] Patwary A A, Fu A, Kumar S and Kumar R 2020 FogAuthChain: A secure location-based authentication scheme in fog computing environments using Blockchain *Comput. Commun.* 162 212–24.
- [58] Jaikla T, Vorakulpipat C, Rattanalerdnorsorn E and Hai H D 2019 A secure network architecture for heterogeneous IoT devices using role-based access control 2019 27th Int. Conf. Software, Telecommun. *Comput. Networks, SoftCOM* 2019.
- [59] Mistry I, Tanwar S, Tyagi S and Kumar N 2020 Blockchain for 5G-enabled IoT for industrial automation: A systematic review, solutions, and challenges *Mech. Syst. Signal Process.* 135 106382.
- [60] Song L, Zhu Z, Li M, Ma L and Ju X 2021 A Novel Access Control for Internet of Things Based on Blockchain Smart Contract *IEEE Adv. Inf. Technol. Electron. Autom. Control Conf.* 2021 111–7.
- [61] Chiu W Y, Meng W and Jensen C D 2021 My data, my control: A secure data sharing and access scheme over blockchain *J. Inf. Secur. Appl.* 63 103020.
- [62] Cruz J P, Kaji Y and Yanai N 2018 RBAC-SC: Role-based access control using smart contract *IEEE Access* 6 12240–51.
- [63] Wan Muhamad W N, Matrazali N, Ishak K, Hasbullah N, Zainudin N, Ramli S, Wook M, Ishak Z and MSaad N 2019 Enhance Multi-factor Authentication Model for Intelligence Community Access to Critical Surveillance Data pp 560–9.
- [64] Putra G D, Dedeoglu V, Kanhere S S, Jurdak R and Ignjatovic A 2021 Trust-Based Blockchain Authorization for IoT *IEEE Trans. Netw. Serv. Manag.* 18 1646–58.
- [65] Liu Y, Lu Q, Chen S, Qu Q, O'Connor H, Raymond Choo K K and Zhang H 2020 Capability-based IoT access control using blockchain *Digit. Commun. Networks* 0–6.
- [66] Pal S, Rabehaja T, Hitchens M, Varadharajan V and Hill A 2020 On the Design of a Flexible Delegation Model for the Internet of Things Using Blockchain *IEEE Trans. Ind. Informatics* 16 3521–30.
- [67] Ali G, Ahmad N, Cao Y, Ali Q E, Azim F and Cruickshank H 2019 BCON: Blockchain based access CONtrol across multiple conflict of interest domains *J. Netw. Comput. Appl.* 147 102440.
- [68] Abou-Nassar E M, Iliyasa A M, El-Kafrawy P M, Song O Y, Bashir A K and El-Latif A A A 2020 DITrust Chain: Towards Blockchain-Based Trust Models for Sustainable Healthcare IoT Systems *IEEE Access* 8 111223–38.

- [69] Taherkordi A and Herrmann P 2018 Pervasive Smart Contracts for Blockchains in IoT Systems 6–11.
- [70] Hakiri A and Dezfouli B 2021 Towards a Blockchain-SDN Architecture for Secure and Trustworthy 5G Massive IoT Networks SDN-NFV Sec 2021 - Proc. 2021 ACM Int. Work. Softw. Defin. Networks Netw. Funct. Virtualization Secur. co-located with CODAYSPY 2021 11–8.
- [71] Egala B S, Pradhan A K, Badarla V and Mohanty S P 2021 Fortified-Chain: A Blockchain-Based Framework for Security and Privacy-Assured Internet of Medical Things with Effective Access Control IEEE Internet Things J. 8 11717–31.
- [72] Sanwar Hosen A S M, Singh S, Sharma P K, Ghosh U, Wang J, Ra I H and Cho G H 2020 Blockchain-Based Transaction Validation Protocol for a Secure Distributed IoT Network IEEE Access 8 117266–77.
- [73] Razali N A M, Muhamad W N W, Ishak K K, Saad N J A M, Wook M and Ramli S 2021 Secure Blockchain-Based Data-Sharing Model and Adoption among Intelligence Communities IAENG Int. J. Comput. Sci. 48.
- [74] Guruprakash J and Koppu S 2020 EC-ElGamal and Genetic Algorithm-Based Enhancement for Lightweight Scalable Blockchain in IoT Domain IEEE Access 8 141269–81.
- [75] Xu L, Chen L, Gao Z, Fan X and Shi W 2020 DL-DP: Improving the security of industrial IoT with decentralized ledger defined perimeter BSCI 2020 - Proc. 2nd ACM Int. Symp. Blockchain Secur. Crit. Infrastructure, Co-located with AsiaCCS 2020 53–62.
- [76] Syed T A, Siddique M S, Nadeem A, Alzahrani A, Jan S and Khattak M A K 2020 A Novel Blockchain-Based Framework for Vehicle Life Cycle Tracking: An End-to-End Solution IEEE Access 8 111042–63.
- [77] Pajooh H H, Rashid M, Alam F and Demidenko S 2021 Hyperledger fabric blockchain for securing the edge internet of things Sensors (Switzerland) 21 1–29.
- [78] Serrano W 2021 The Blockchain Random Neural Network for cybersecure IoT and 5G infrastructure in Smart Cities J. Netw. Comput. Appl. 175 102909.
- [79] Tan L, Shi N, Yu K, Aloqaily M and Jararweh Y 2021 A Blockchain-empowered Access Control Framework for Smart Devices in Green Internet of Things ACM Trans. Internet Technol. 21.
- [80] Pyoung C K and Baek S J 2020 Blockchain of Finite-Lifetime Blocks with Applications to Edge-Based IoTa IEEE Internet Things J. 7 2102–16.
- [81] Ali G, Ahmad N, Cao Y U E, Khan S, Cruickshank H, Qazi E A L I and Ali A 2020 xDBAuth: Blockchain Based Cross Domain Authentication and Authorization Framework for Internet of Things 8.

# Research on Intelligent Control System of Air Conditioning based on Internet of Things Intelligent Control System of Air Conditioning

Binfang Zhang

Department of Art and Design  
Shijiazhuang University of Applied Technology  
Shijiazhuang, 050000, China

**Abstract**—The current air conditioning intelligent control system cannot achieve the ideal energy-saving effect. The indoor temperature and humidity control is not good enough either. Therefore, an intelligent air-conditioning control system based on Internet of Things technology is designed. The hardware part of the system includes system control motherboard, sensor module, execution control structure, wireless communication module and access layer. The software includes the design of communication layer, the design of monitoring management, and the design of intelligent indoor air-conditioning temperature remote control algorithm. The experimental results show that the control effect of the intelligent air conditioner is more accurate and energy-saving, the opening degree of the air conditioning valve is larger, and the comfort is improved. The indoor temperature and humidity of the proposed system are both more ideal.

**Keywords**—Internet of things technology; intelligent control of air conditioning; system design; double closed-loop load; virtual synchronizer of air conditioning

## I. INTRODUCTION

In recent years, science and technology have developed rapidly, and smart home control has become the most important part of the development of the decoration field. The research and development of smart home can greatly reduce the energy consumption of electrical equipment, reduce the cost of manual operation and management, and provide customers with more comfort and humane environment [1-2]. In the intelligent building, the central air-conditioning system consumes a lot of electricity, and it is the most important part in building the automation control system. In modern buildings, warm air conditioning is the basic supporting equipment, which can adjust the overall temperature and humidity of the building. Heating air conditioning system accounts for more than half of the total electricity consumption of the whole building, so it is necessary to find a suitable technology to reduce its energy consumption. Relevant scholars have found that intelligent control is the core of the entire intelligent system. In order to achieve low power consumption and low cost, and to ensure the safe and stable operation of the air conditioning system, reliable algorithms and selection of hardware and software are required to ensure the function of the system. [3-4].

However, at present, most of the air-conditioning intelligent control systems of engineering projects are idle and

resources are wasted. This is due to the design mistakes of relevant managers and the inability of air-conditioning intelligent systems to meet the needs of sustainable development and cost control. Researchers in this field have conducted relevant research on the above issues. For example, reference Yan Junwei et al. proposes an energy consumption prediction method based on machine learning divided operation modes. Firstly, the k-means algorithm is used to divide the system operation mode, and the main factors affecting the operation mode are selected by the random forest method. Then, the prediction model is established by BP neural network to predict the mode and energy consumption in turn [5]. Fu Huansen et al. designs a working algorithm of the air conditioning control system in different seasonal modes, analyzes the key points of PID control in Siemens s7-1200plc programming, and takes the combined air conditioning of an actual engineering project of a pharmaceutical group as an example to design the function of the touch screen configuration interface. The system ensures the priority control of humidity in the drug warehouse during the program design [6]. Li Xiaotong et al aiming at the characteristics of nonlinearity, large delay of central air conditioning system and the difficulty of establishing accurate model, proposes a temperature control method of central air conditioning system based on model free reinforcement learning. Aiming at the communication problem between Energy PI us and MATLAB, MLE + tool is used to realize the joint simulation of them. Comparing reinforcement learning algorithm with benchmark start stop strategy and model predictive control strategy, the method can minimize the energy consumption of air conditioning system on the premise of ensuring comfort [7].

According to the above-mentioned current domestic and foreign researches, the existing air-conditioning control effects cannot take into account low power consumption and intelligent effects, and most of the air-conditioning valves have a low opening degree, which makes it difficult to bring users a more comfortable and humanized experience. In order to solve the above problems, the research designs dual closed-loop loads and perfect communication layer, monitoring management and remote-control functions of intelligent air conditioners on the basis of the Internet of Things with intelligent improvement strategies.



## II. DESIGN OF AIR CONDITIONING INTELLIGENT CONTROL SYSTEM BASED ON INTERNET OF THINGS TECHNOLOGY

From the composition point of view, the air-conditioning remote control system is mainly composed of wireless controller, mobile phone software, remote management and data analysis system. With the help of the sensor, the communication module of wireless controller can get the air conditioning running status information in time. Through the wireless network, the user can set the temperature in advance. Far away, users can use mobile phones or computer software to operate the system. According to the basic operation condition, the management system can send the corresponding control command. After receiving the command, the controller can control the air-conditioning system in real time by sending infrared coded signal.

### A. Application Principle of Internet of Things in Air Conditioning Control

The rapid development of Internet of Things has provided an important basis for the application of smart house. In the application of the Internet of Things, we should take full advantage of wireless technology, such as radio frequency, Bluetooth, infrared sensing, ZigBee and WIFI. ZigBee technology's energy consumption is low, and the cost is less than other technologies. It can be widely promoted in the future development. Wireless sensor networks are mainly deployed in the monitoring area, which is a self-organizing multi-hop form. There are a large number of sensor nodes, using the characteristics of sensor cheap [8-9]. After the application program and physical setup are completed, the user is authorized by the Web or application program, and then the command to be executed can be sent out in a graphical way. After the central controller receives the command, it requires the simulative starter to perform the corresponding operation. When the user doesn't give the operation instruction, the central controller can use the network technology and wireless sensor to receive the change of the outside environment, and then judge the safety factor effectively. When there is a danger, the central controller will find the corresponding security warning, and the central control will activate the scheduled function to notify or alarm the user [10]. When the external environment is relatively safe, the central controller will activate the corresponding control information when it exceeds the set range.

### B. System Hardware Design

#### 1) System Control Motherboard

The main control board of the system adopts the Arduino Mega2560 SCM development board with USB interface, with 54 digital input/output (16 of which can be used as PWM output), 16 analog input, 4 UART interfaces, 1 16MHz crystal oscillator, 1 USB port, 1 power socket, 1 ICSP header and 1 reset button. The motherboard is easy to use, can meet the requirements of air-conditioning control system, and has a more stable performance.

#### 2) Sensor Module

The system sensor module includes temperature and humidity sensor, HC-SR501 human body sensor and photoresistor. Two temperature sensors are respectively installed in the air outlet and the side of the air conditioner. The difference of temperature data is used to judge the on-off state and the refrigeration and heating state. HC-SR501 human body sensor is an automatic control module based on infrared technology. It adopts imported LHI778 probe from Germany, with high sensitivity and strong reliability. HC-SR 50 can be used to judge whether the user is near the air conditioner or not.

#### 3) Enforcement Control Structure

In the remote-control system, the infrared ray sends the instruction information. After receiving the infrared code, the remote management carries on the next instruction operation according to the information characteristic. Because there are many air-conditioning manufacturers, there are some differences in infrared protocol. During the test, the infrared information of different remote controllers should be collected and analyzed. Analyzing the infrared signals of different brands of air-conditioning and storing them according to the requirements, so as to improve the expansibility of the system and make it can be used in different air-conditioning models.

#### 4) Wireless Communication Module

As shown in Fig. 1, the system adopts ESP8266 serial port WIFI chip, which is a complete and self-contained WIFI network solution specially developed for wireless connection requirements. Temperature and humidity sensor can timely obtain indoor temperature and humidity information, through collecting and uploading information, so as to monitor indoor temperature and humidity in real time and achieve reasonable control of temperature and humidity.

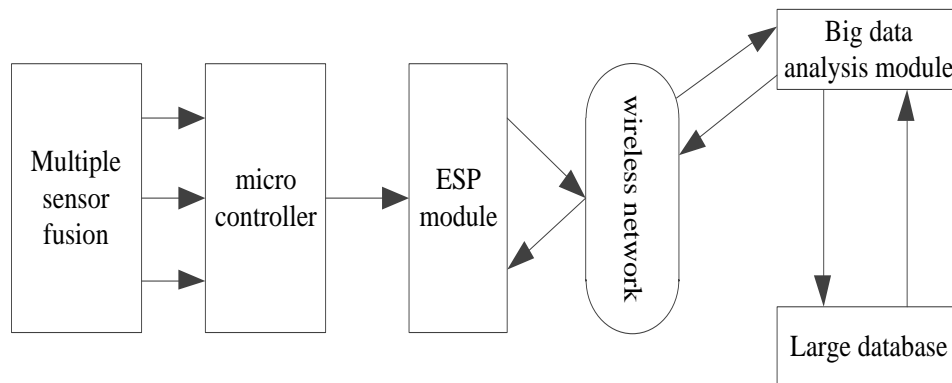


Fig. 1. System Framework

### 5) Access Layer

Wireless sensor network constitutes a perceptual control layer. It collects data information, uploads these data on the basis of network technology, and transfers them to the service management. Intelligent gateway plays the role of access layer, which can integrate multiple protocols, transfer the corresponding data by ZigBee protocol, and achieve the goal of Internet communication. The intelligent network also has the storage function. Through uploading the newest data, it helps the user to use the intelligent gateway to realize the information inquiry the function.

#### a) Control Architecture Design

As shown in Fig. 2, the basic control architecture of the intelligent control system for air conditioning. As a direct power supply device of air-conditioning virtual synchronizer,

storage battery can provide AC electronic supply with voltage between 220-380V. AC suction nozzle is a standard AC device, which can provide AC current with frequency between 360-800HZ for air-conditioning virtual synchronizer. The main control architecture consists of a virtual cabinet and a synchronous load cabinet. Virtual cabinets contain air conditioning temperature display, temperature control chassis, virtual operators, TUR equipment and a part of the blank reserved area. Under the condition of long-term load operation, there will be a large number of control instructions in the system, which are stored in the blank reserved area until the transmission channel is idle. Synchronous load chassis includes 429 bus module, load frequency conversion module, air conditioning power analog output module, load voltage acquisition module. Each module has different physical execution function.

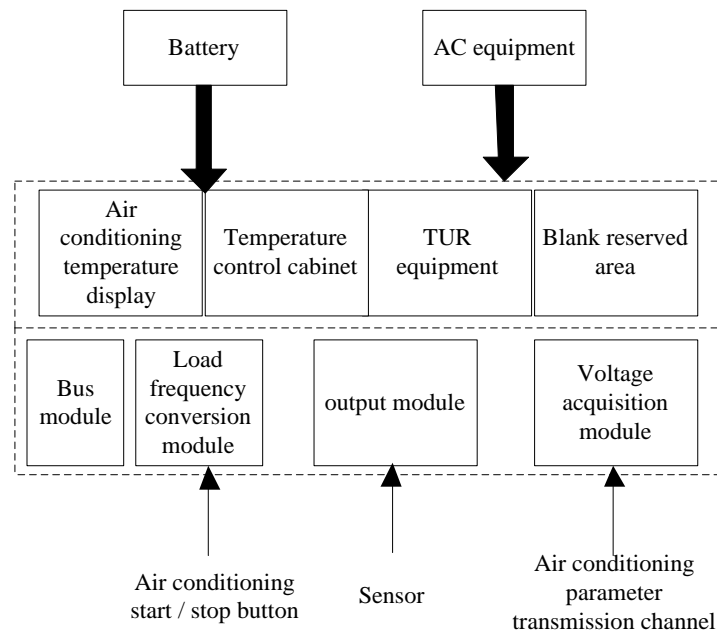


Fig. 2. System Hardware Control Architecture.

#### b) Design of Governor for Virtual Synchronous Machine of Air Conditioner

The traditional DC/AC inverter topology is used to adjust the output power of storage battery by collecting the modal information of load electronic power supply. From the functional point of view, the virtual synchrotron governor is the key subordinate structure of the control architecture, which can integrate the fixed virtual load electronics in the original motor, and provide a certain excitation effect for these electronics with the support of the generator. When the

air-conditioned power analog output module sends out enough asynchronous sensing signals, the virtual power grid of the synchrotron changes from a closed state to an open state, absorbs all the load electronics, and temporarily stores them in the virtual control chip. In order to ensure the reasonable distribution of load electrons, two power-consuming components with the same resistance as the load voltage acquisition module must be installed around the generator, and the internal operating current of the governor must always be rated control current. The complete structure of the virtual synchronizer governor is shown in Fig. 3.

c) Load Double Closed Loop Design

The load double closed-loop design includes two parts: synchronous controller and virtual load unit. According to the regulation rules and control requirements of air-conditioning equipment, the synchronous controller can control the power load signal unilaterally, and the virtual load unit can obtain more AC load signals by transmitting the power consumption signal to the central processor of air-conditioning. In the case of higher requirements for system control instructions, the load double closed-loop system can choose parallel or series operation mode according to the power output of the system to ensure that the air-conditioning virtual synchronizer governor always has sufficient AC electrons [11-12]. In parallel mode, the load double closed-loop, air-conditioning virtual synchronous governor can be selectively connected to the control circuit at the same time, effectively avoiding the arbitrary notice of the battery system. In series mode, the load double closed-loop, air-conditioning virtual synchronizer governor can also be connected to the control circuit at the same time, but in this case, the automatic load electrons can easily reach the maximum value. Fig. 4 shows the detailed load double closed loop structure.

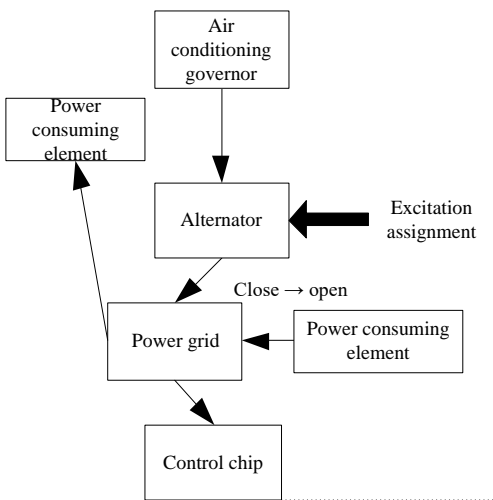


Fig. 3. Structure Diagram of Air Conditioner Virtual Synchronous Machine Governor.

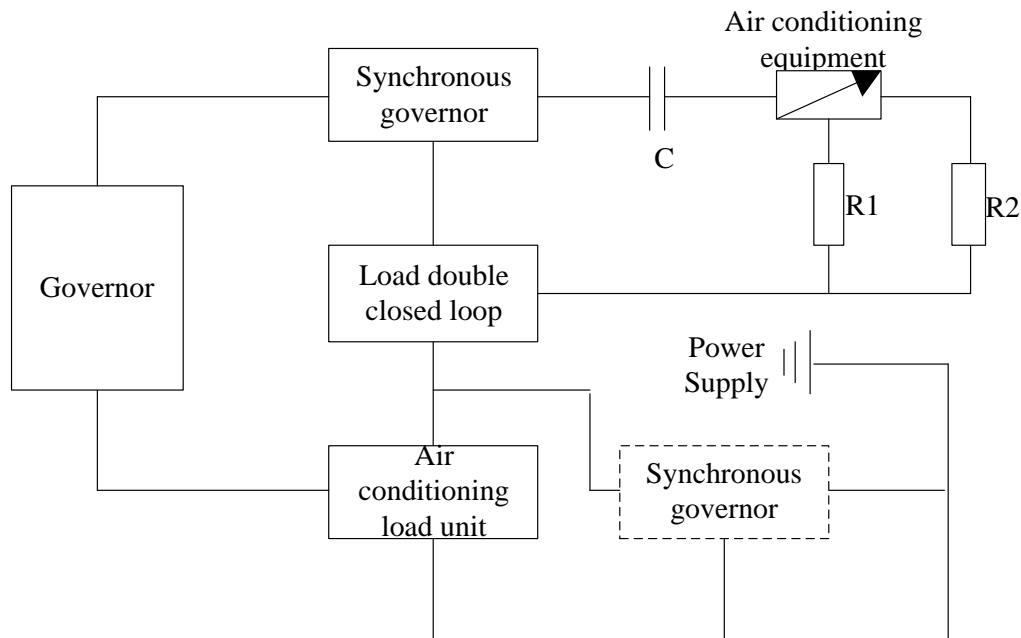


Fig. 4. Load Double Closed Loop Structure.

C. System Software Design

1) Communication Layer Design

The Modbus communication protocol in the communication layer adopts RTU transmission mode to realize data reading. RTU transmission mode message frame format: function code is 8, CRC check code is 16 bits, data area is  $n * 8$  bits, and slave bit address is 8. See Table I for common command codes and functions in the system of Modbus communication protocol.

TABLE I. COMMON COMMAND CODES AND FUNCTIONS IN THE SYSTEM OF MODBUS COMMUNICATION PROTOCOL

Command code	Effect
01	Read controller value
02	Forced input of single switching value
03	Read switch status
04	Force input controller value
05	For exception response
06	Communication diagnosis
07	On off state

### 2) Design of Monitoring Management

Through monitoring the management computer remote monitoring software with ModBus communication protocol and PLC communication, Real-time on-line monitoring of the operation of HVAC system is realized. The steps for remote online monitoring are shown in Fig. 5.

In order to convert the remote intelligent control mode of the system, the control signal is transmitted to the PCL control layer through the command code of ModBus communication protocol in remote online monitoring software. If the

automatic control mode is selected, the PCL selects the data in the controller, operates on the collected signal of HVAC according to the fuzzy self- adaptive PID controller. It judges the operation result, and automatically controls the stop or start of HVAC equipment. If you choose the manual control mode, in order to control the HVAC equipment, the input of single equipment switch is realized by ModBus communication protocol, the switch of relay in PLC is changed, and the manual remote control of HVAC equipment is stopped or started.

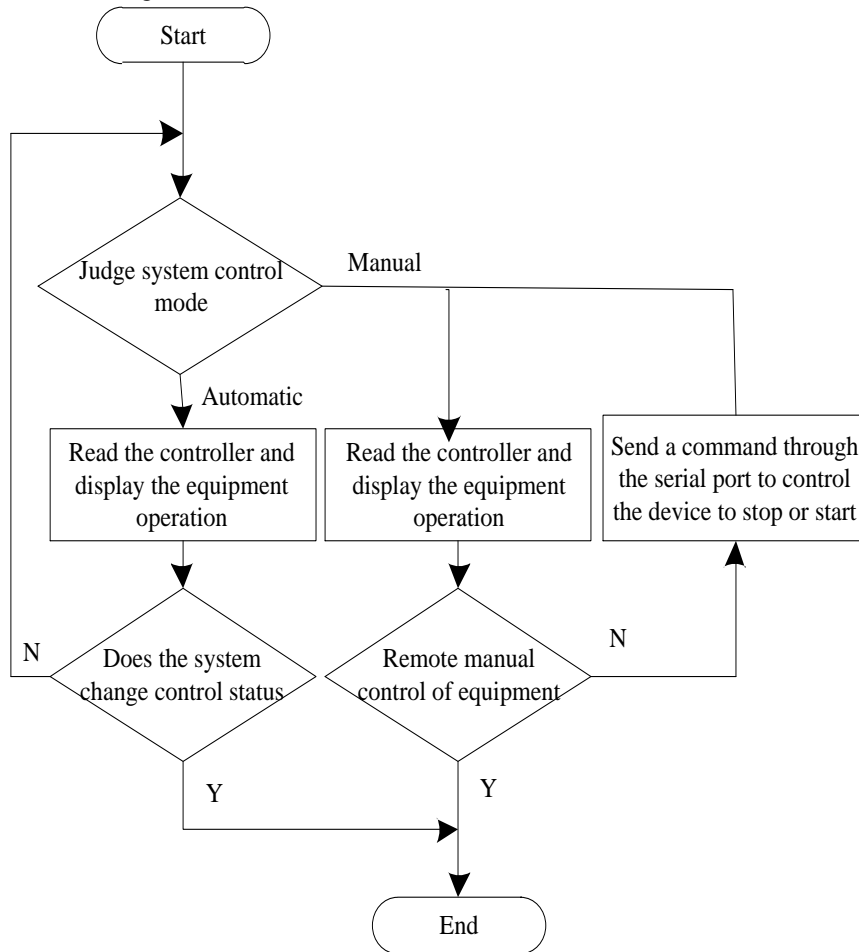


Fig. 5. Steps of Remote Online Monitoring.

### 3) Implementation of Remote-Control Algorithm for Intelligent Indoor Air Conditioning Temperature

The control structure of influencing factors of remote intelligent indoor air conditioning temperature control is shown in Fig. 6. The data actually measured by different sensors are sent to the PID controller after being adjusted, and the selection signal generated based on the parameters optimized by the PID controller, such as  $k_p$ ,  $k_i$  and  $k_d$ . They are used to keep the indoor air conditioner running and keep the room temperature, supply air temperature and supply air volume within the error control range [13-14].

The energy-saving principle of HVAC uses frequency

conversion technology to realize frequency conversion control of temperature, but frequency conversion control is unable to build accurate mathematical model. So, the fuzzy adaptive PID controller is added into the PLC control layer. In order to improve the control quality and make the PID controller have the intelligent performance of fuzzy control, the fuzzy adaptive PID controller is generated by combining the conventional PID control and fuzzy control. Fuzzy adaptive PID control principle, see Fig. 7.

The parameters of each variable, fuzzy control rule, fuzzy subset and the membership function of input and output variables are designed by fuzzy adaptive PID controller. Initial setup of fuzzy adaptive PID controller, see Fig. 8.

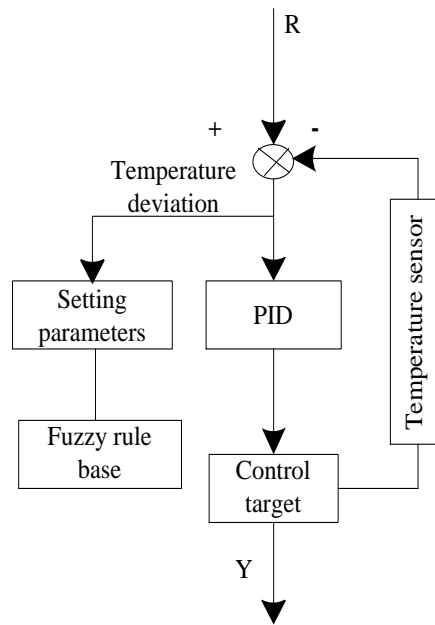


Fig. 6. Principle of Fuzzy Adaptive PID Control.

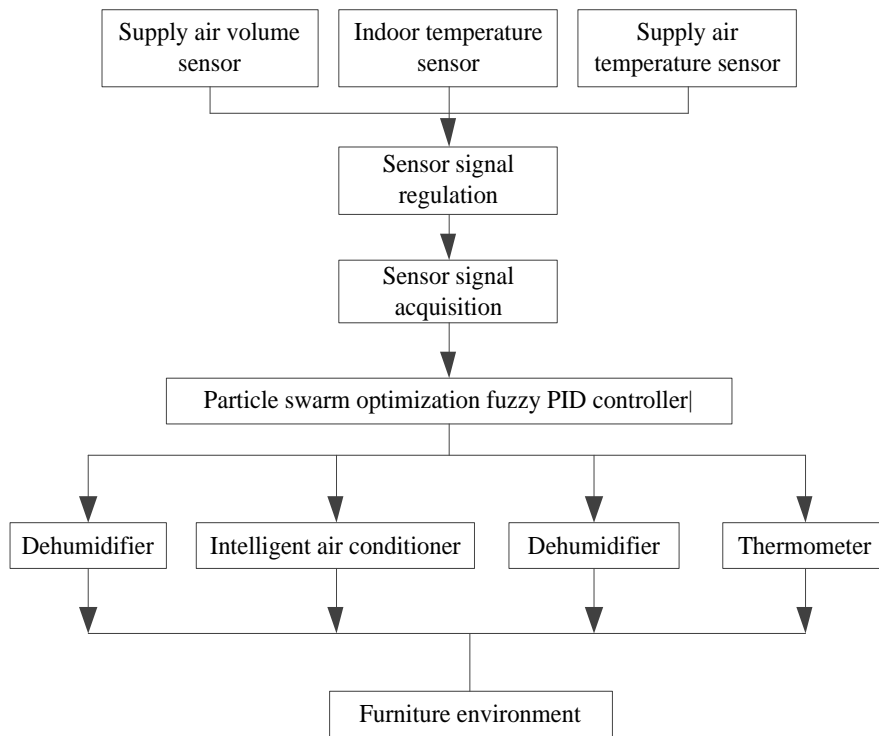


Fig. 7. Control Structure of Room Temperature, Air Supply Temperature and Air Supply Volume.

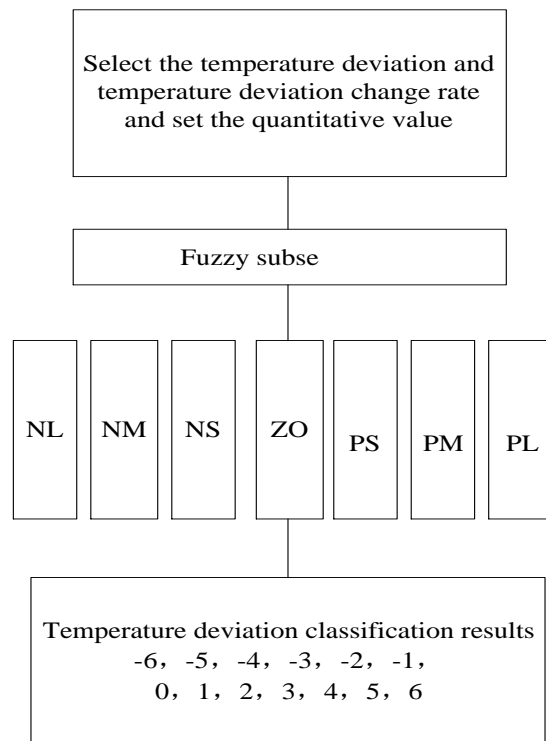


Fig. 8. Initial Setting of Fuzzy Adaptive PID Controller.

If  $K_d$  and  $K_p$  change in the predetermined range of  $[K_{d,\min}, K_{d,\max}]$  and  $[K_{p,\min}, K_{p,\max}]$ , normalized  $K_d$  and  $K_p$  convert them to the parameters between 0 and 1. The newly converted parameters are described by  $K_d'$  and  $K_p'$  respectively, and their conversion process is described by formulas (1) and (2) respectively:

$$K_d' = \frac{K_d - K_{d,\min}}{K_{d,\max} - K_{d,\min}} \quad (1)$$

$$K_p' = \frac{K_p - K_{p,\min}}{K_{p,\max} - K_{p,\min}} \quad (2)$$

Formula (3) represents the relationship between integral and differential time constants:

$$T_i = \beta T_d \quad (3)$$

Where  $T_d$  represents the differential time constant,  $\beta$  represents the parameter, and  $T_i$  represents the integral time constant.

The integral gain solution is described by formula (4):

$$K_i = \frac{K_p}{\beta T_d} = \frac{K_p^2}{\beta K_d} \quad (4)$$

The membership function of temperature deviation and temperature deviation change rate is a fuzzy subset language

variable, which is described by  $\{NL, NM, NS, ZO, PS, PM, PL\}$ , and each value width is consistent and distributed in a triangle.

In order to judge the reasoning rules of parameters  $K_d'$ ,  $K_p'$  and  $\beta$ , the decision is made according to the temperature deviation, the knowledge base summarized by experts (fuzzy rule base) and the change rate of temperature deviation. The rules are as follows:

if  $e(k)$  is  $A_i$ ,  $\Delta e(k)$  is  $B_i$ , then  $K_p'$  is  $C_i$ ,  $K_d'$  is  $D_i$ ,  $\beta = \beta_i$  ( $i = 1, 2, \dots, m$ ).

Among them,  $A_i$ ,  $B_i$ ,  $C_i$ ,  $D_i$  represent the language variable of  $e(k)$ ,  $\Delta e(k)$ ,  $K_p'$ ,  $K_d'$ ,  $\beta$ , a fuzzy relation has a rule, 7 fuzzy subsets have 49 fuzzy control rules [15]. Fuzzy control table is made up of 49 Fuzzy rules stored by computer. Fuzzy control table is made up of the relation between each input and output parameter according to expert's knowledge and experience, and then the Fuzzy control rules are made.

### III. EXPERIMENTAL DESIGN AND RESULT ANALYSIS

In order to verify the effectiveness of the intelligent control system based on Internet of Things, a simulation experiment is designed. The experimental comparison method is the composite air-conditioning control system and the central air-conditioning control system based on reinforcement learning proposed in reference [6] and [7] respectively. The PID parameters are set to  $k_{p0} = 0.18$ ,  $k_{i0} = 0.0012$  and



$k_{d0} = 0.5$ , and the input quantization factors of the PID controller are set to  $k_{up} = 0.005$ ,  $k_{ui} = 0.02$  and  $k_{ud} = 0.5$ . After setting the relevant values of each characteristic

parameter, the traditional PID algorithm and the PID control algorithm based on particle swarm optimization are analyzed, and the corresponding curve of the valve opening of each end of the PID controller is obtained as shown in Fig. 9.

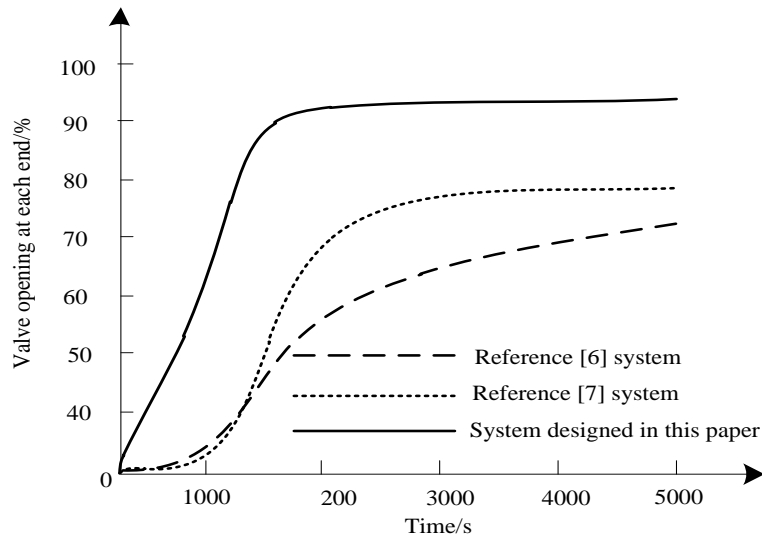


Fig. 9. Valve Opening at Each End of Household Intelligent Air Conditioning Structure.

As shown in Fig. 9, the operation time of the air-conditioning structure is expressed in abscissa (unit: s); the change in the opening of the air-conditioning valve at each end of the air-conditioning structure is expressed in ordinate. The results show that the opening trend curve is stable and the dynamic regulation performance is more stable than that of the traditional control method, and the opening trend fluctuates greatly and the period is longer. When the air-conditioning structure works normally, the opening degree of the traditional air-conditioning valve is about 78%, and the opening degree of the air-conditioning valve obtained by this method can be controlled at 95%, so the friction loss formed by the air-conditioning valve in the actual work of the air-conditioning structure can be reduced to the greatest extent. It is proved that the intelligent air-conditioning structure of modern home remote control realized by this method is more accurate and energy-saving. The larger the opening degree of the air-conditioning valve is, the lower the operating noise of the corresponding air-conditioning structure is, and the comfort degree of the operating environment of the air-conditioning structure for users is improved.

Set the output variable is the indoor temperature, the input variable is the frequency of HVAC compressor. Transfer function model of setting room temperature and frequency of HVAC compressor, described by formula (5):

$$H(s) = \frac{G'}{Ts + 1} e^{-\phi s} \quad (5)$$

Where,  $s$  represents the complex variable,  $T$  represents the system time constant,  $G'$  represents the open-loop gain,

which is set to  $0.6^\circ\text{C} / \text{Hz}$ , and  $D$  represents the time constant of the delay link, which is set to 1000s.

In order to verify the effectiveness of the system in this paper, the system test is carried out with MATLAB. The comparison systems used in the experiment are the combined air conditioning control system proposed in reference [6] and the central air conditioning control system based on reinforcement learning proposed in reference [7]. The upper limit frequency of HVAC compressor is 240Hz, the indoor initial temperature is  $0^\circ\text{C}$ , the set temperature is  $24^\circ\text{C}$ , and the step input is applied to the three experimental systems. At the same time, the set temperature drops to  $20^\circ\text{C}$  at 5000ms and rises to  $24^\circ\text{C}$  at 9000s. After repeated tests, the indoor temperature response time curves of the three systems are recorded and described in Figure 10.

As can be seen from Fig. 10, the room temperature response speed and temperature control of this system are obviously higher than those of the other two systems. This system can quickly control the air conditioner and make the room temperature reach the set temperature. At 5000s, the room temperature drops to  $22^\circ\text{C}$ , and at 9000s, the room temperature rises to  $24^\circ\text{C}$  with a stable trend. Among them, the room temperature of document [7] system falls to  $22^\circ\text{C}$  only at 7000s, and the document [6] system is always in a fluctuating state, which does not meet the requirements of the set temperature, indicating that this system has a better temperature control effect.

In the cooling experiment, when the delay is 1000s, three kinds of system cooling-delay time curves are depicted in Fig. 11.

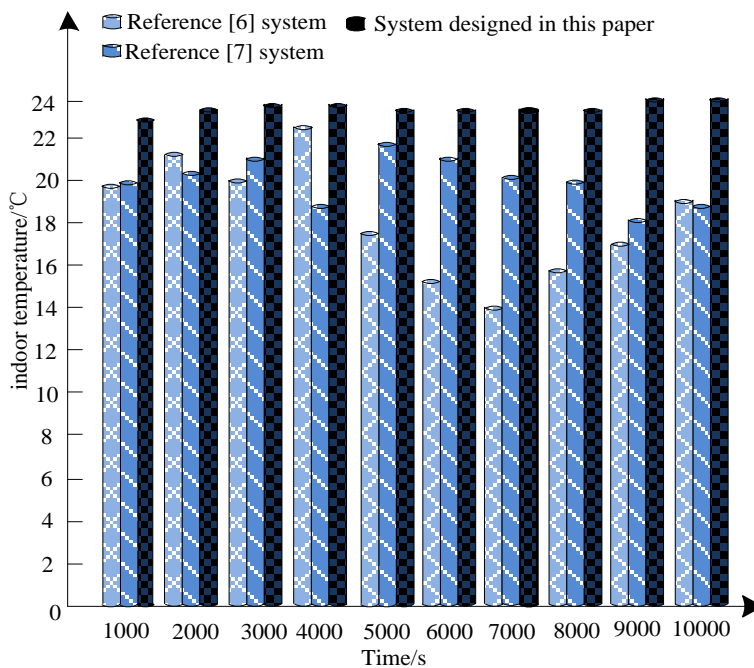


Fig. 10. Indoor Temperature Response Curve.

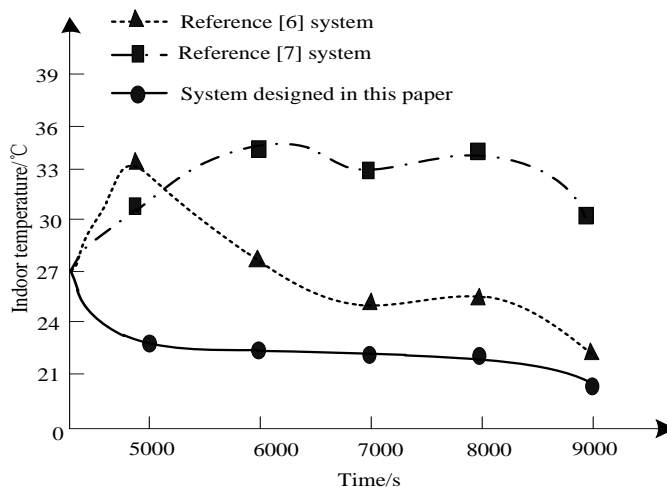


Fig. 11. Cooling Delay Time Curve.

It can be seen from Fig.11 that with the increase of the delay time, the cooling time curve of the system in this paper has little fluctuation and the cooling time is fast, while the cooling time curve of the other two systems has great fluctuation and the cooling time is slow, which shows that the cooling performance of the system in this paper is good.

In order to improve the energy-saving and emission reduction of HVAC, indoor relative humidity was set at 40% and CO<sub>2</sub> content was set at 70% in the experiment.

Fig. 12 shows that the effect of indoor relative humidity control of this system is superior to that of other two systems. The control effect of this system is less than 30%, the other two systems are not good, and the fluctuations of indoor

relative humidity are large.

Fig. 13 shows that the control effect of CO<sub>2</sub> volume fraction of this system is obviously better than that of the other two systems. The indoor CO<sub>2</sub> content of this system reaches 3000 ppm at 1000s, after which the control is relatively stable. Although the control time of the system [7] is short and unstable at 4000s, the control time of the system [6] is short and unstable. The system [6] has not met the requirements of the indoor CO<sub>2</sub> content. Therefore, the control of the indoor CO<sub>2</sub> content of this system meets the requirements of energy saving and emission reduction.

Under the condition of 0.77 air conditioning load parameter, the change of load spike frequency is recorded in

60s as the experimental time, respectively, after the application of the control system of the experimental group and the control group (the combined air conditioning control

system proposed in the reference document [6]). Experimental details are shown in Fig. 14.

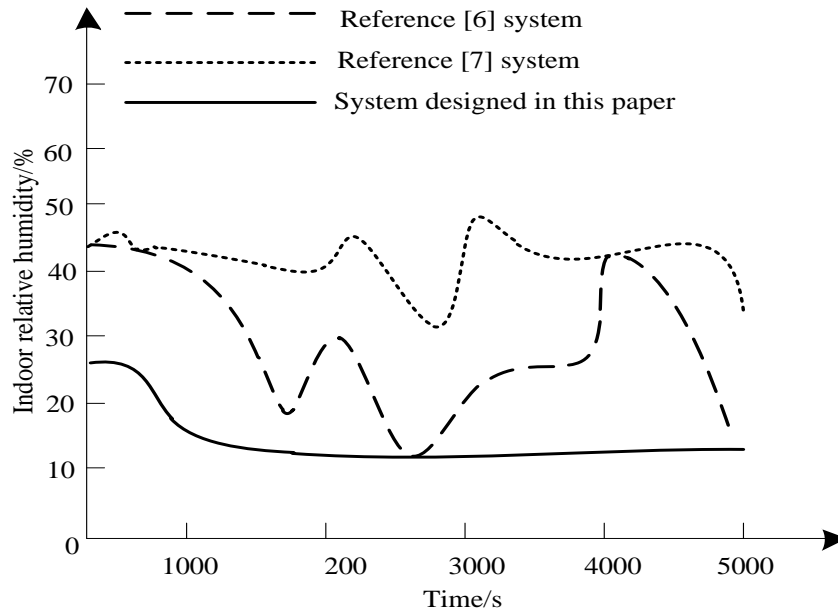


Fig. 12. Indoor Relative Humidity Control Curve.

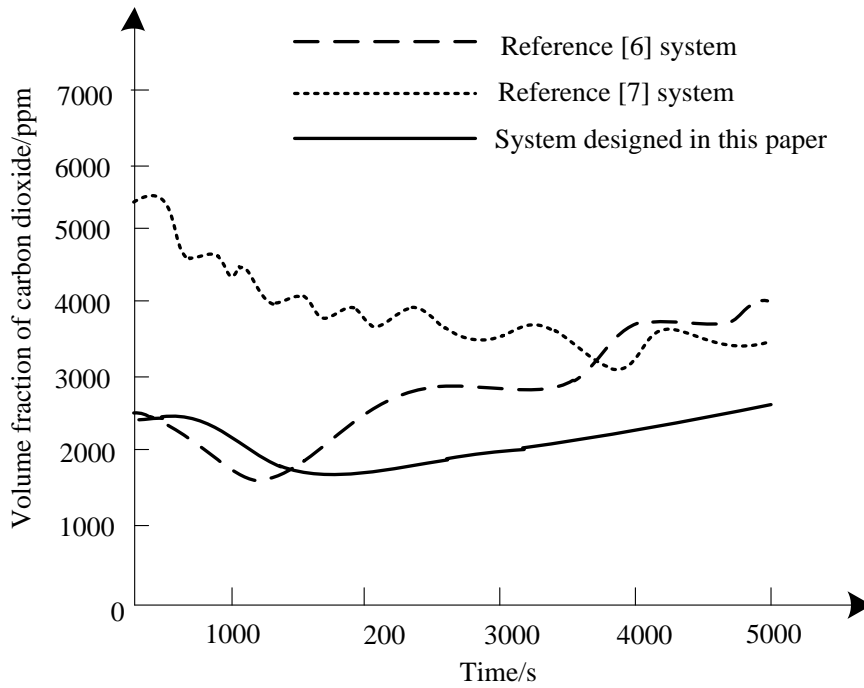


Fig. 13. Indoor Carbon Dioxide Content Control Curve.

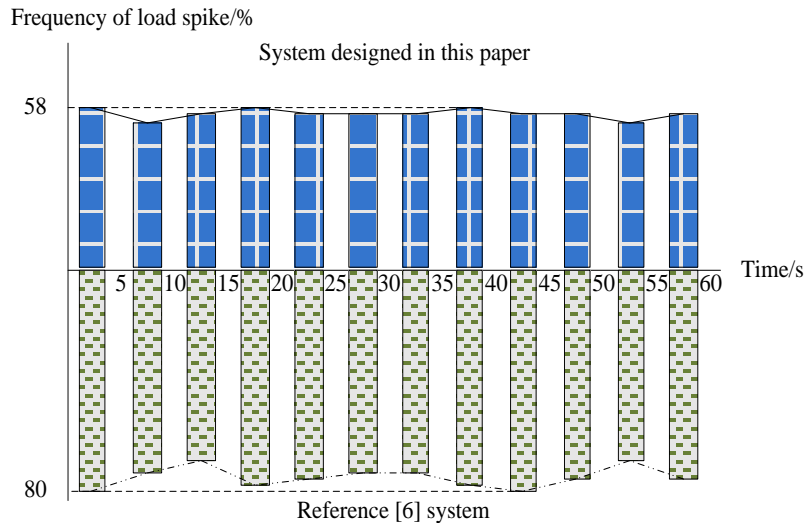


Fig. 14. Frequency Comparison of Load Spikes.

Fig.14 shows that with the increase of experimental time, the frequency of load spike of experimental group presents a more stable trend, the frequency of load spike reaches 58% of the maximum value. The peak frequency of load in the control group also showed a more stable trend, but compared with the experimental group, this trend was significantly worse. The peak frequency of load in the control group reached 80%, much higher than the experimental group. There was no significant change in the peak frequency of load in the control group in the early and late experiment. In conclusion, under the condition of air conditioning load parameter of 0.77, the frequency of load spike can be greatly reduced by using the designed control system.

#### IV. CONCLUSION

In order to solve the current air conditioning, intelligent control system indoor temperature and humidity control effect is not enough ideal problem. Therefore, an intelligent air-conditioning control system based on Internet of Things technology is designed. The application principle of Internet of things in air conditioning control is analyzed. Based on this, the hardware and software of the system are designed. The hardware part of the system includes system control motherboard, sensor module, execution control structure, wireless communication module and access layer. The control architecture of the system and the governor of the virtual synchronizer are designed. In order to realize the energy-saving control of air conditioning, double closed-loop load is designed. The load double closed-loop design includes two parts: synchronous controller and virtual load unit. The software of the system includes the design of communication layer, the design of monitoring management, and the design of intelligent indoor air-conditioning temperature remote control algorithm. The experimental results show that when the air-conditioning structure is working normally, the opening of the traditional air-conditioning valve is about 78%, and the opening of the intelligent control system proposed by the study can be controlled at about 95%. In the research system,

the room temperature decreased from 5000s to 22° C, and the room temperature increased to 24° C in 9 000s. The trend was stable at this stage, but the room temperature of the system in [7] only dropped to 22°C at 7000s, and the system in [6] was always in a state of fluctuation. The control effect of this system is stable at 30%, while the indoor relative humidity of the other two systems fluctuates greatly. The indoor CO<sub>2</sub> content of the system reaches 3000ppm in 1000s, and then the control is in a stable state. The control of indoor CO<sub>2</sub> content can meet the policy requirements of energy conservation and emission reduction. Under the condition that the air-conditioning load parameter is 0.77, the load peak frequency can be greatly reduced by using the designed control system. In summary, the air-conditioning intelligent control system proposed in the study can effectively reduce energy consumption and carbon emissions, and provide users with a more intelligent and comfortable indoor environment. However, there are still shortcomings in the research. With the development of information technology, the application of low-power wide-area network technology in air-conditioning intelligent control systems can be further explored in the future.

#### REFERENCE

- [1] U. S. Dubey, "Design & Cost Estimation of HVAC System for School Building". SSRN Electronic Journal, 2020, 7(2):38-48.
- [2] Y. N. Harmath, "Dynamic B-SIM application for energy consumption estimation of DOAS with FCU HVAC system during design phase". E3S Web of Conferences, 2019, 91(2):02001.
- [3] X. Hu, J. N. utaro. "A Priority-Based Control Strategy and Performance Bound for Aggregated HVAC-Based Load Shaping." IEEE Transactions on Smart Grid, 2020, 11(5):4133- 4143.
- [4] W. Tian, C., Lei, "Tian M. Dynamic Prediction of Building HVAC Energy Consumption by Ensemble Learning Approach. 2018 International Conference on Computational Science and Computational Intelligence" (CSCI). 2018, 254-257.
- [5] J. W Yan., Y. Z. Ma, X. Zhou "Hourly Energy Consumption Prediction Methods for Complex Central Air Conditioning Control Systems Based on Operating Mode Division." Building Science, 2020, 36(2):176-182,190.

- [6] H. S. Fu, G. Q. Wang, J. Cao, et al. "Design and Practice of Combined Air Conditioning Control System." *Electric Drive*, 2019, 49(9):41-45.
- [7] X. T. Li, C. G. Cui, N. Yang, et al. "Temperature Control and Energy Saving of HVAC System Based on Reinforcement Learning." *Computer Simulation*, 2021, 38(4):198-202,224.
- [8] K. A. Meerja, P. V. Naidu, S. Kalva, "Price Versus Performance of Big Data Analysis for Cloud Based Internet of Things Networks." *Mobile Networks & Applications*, 2019, 9:1-17.
- [9] A. Aslam, U. Mehmood, M. H. Arshad, et al. "Dye-sensitized solar cells (DSSCs) as a potential photovoltaic technology for the self-powered internet of things (IoTs) applications." *Solar Energy*, 2020, 207:874-892.
- [10] Y. Huo, C. Meng, R. Li., et al. "An overview of privacy preserving schemes for industrial Internet of Things." *China Communications*, 2020, 17(10):1-18.
- [11] J. Yu, Q. Liu, A. Zhao, et al. "Optimal chiller loading in HVAC System Using a Novel Algorithm Based on the distributed framework." *Journal of Building Engineering*, 2019, 28:101044.
- [12] M. Toub, M. Shahbakhti, R. Robinett, et al. "Model Predictive Control of Micro-CSP Integrated into a Building HVAC System for Load Following Demand Response Programs." *ASME 2019 Dynamic Systems and Control Conference*. 2019, 45(2):563-572.
- [13] F. A. Qureshi, C. N. Jones, "Hierarchical Control of Building HVAC System for Ancillary Services Provision." *Energy & Buildings*, 2018, 169(JUN.):216-227.
- [14] C. E. Huang, C. Li, X. Ma, "Active-Disturbance-Rejection-Control for Temperature Control of the HVAC System." *Intelligent Control & Automation*, 2018, 09(1):1-9.
- [15] M. Ning, M. Zaheeruddin, "Neural Network Model-Based Adaptive Control of a VAV-HVAC&R System." *International Journal of Air Conditioning & Refrigeration*, 2019, 27(1):1950006.

# A Short Review on the Role of Various Deep Learning Techniques for Segmenting and Classifying Brain Tumours from MRI Images

Kumari Kavitha. D<sup>1</sup>

Department of Electronics and Communication Engineering  
Koneru Lakshmaiah Education Foundation  
Green Fields, Guntur, India

E. Kiran Kumar<sup>2</sup>

Department of Electronics and Communication Engineering  
Koneru Lakshmaiah Education Foundation  
Green Fields, Guntur, India

**Abstract**—The past few years have observed substantial growth in death rates associated with brain tumors and it is second foremost source of cancer-related demises. However, it is possible to increase the chance of survival if tumors are identified during initial stage by employing various deep learning techniques. These techniques are helpful to the doctors during the diagnosis process. The MRI which refers to magnetic resonance imaging is a non-invasive procedure and low ionization radiation diagnostic tool to evaluate an abnormality that evolves in the form of shape, location or position, size and texture of tumour. This paper focuses on the systematic literature survey of numerous Deep-Learning methods with suitable approaches for tumour segmentation and classification (normal or abnormal) from MRI images. Furthermore, this paper also provides the new aspects of research and clinical solution for brain tumor patients. It incorporates Deep-Learning applications for accurate tumor detection and quantitative investigation of different tumor segmentation techniques.

**Keywords**—Medical image segmentation; convolutional neural networks (CNN); deep-CNN; feed forward neural networks; brain tumor segmentation (BraTS) and U-net

## I. INTRODUCTION

Brain tumors are neurological fatal disorders, a bunch of atypical cells that are found increasing in the human brain or by the surroundings of the brain that which affects the normal brain cells and further results in cancer [1]. They are classified into two variants malignant, which is cancer causing and benign which is non-cancerous. Benign tumours are less offensive, form gradually and are isolated from normal tissues regularly. Malignant tumors develop quickly, lack of defined boundaries, and are difficult to identify from normal tissues [1, 2]. These tumors cause more pain inside the brain and can migrate to the spinal cord. At the same time, malignant tumors are complex to eliminate entirely from the cells in the brain; moreover these tumours have tendency to transform to cancer which is deadly. The second prime reason for most cancer related deaths is the malignant kind of brain tumour [3].

As per statistics, there are approximately 12.7 million cancerous persons in the world each year, with 7.6 million people dying because of cancer [4]. The Hindu published an article in 2016 that reported that around 2,500 children of India suffered from malignant tumors every year and every year 20% of children were diagnosed with brain tumors if 4,000 to 5000

people were diagnosed [4, 5]. Conferring to the American Brain Tumor Association (ABTA), about 78,000 new brain tumour cases will be diagnosed by year end of 2018. According to a poll conducted by the “Times of India,” approximately three million individuals in India suffer from cancer, with one million being diagnosed with new kinds of cancer. As stated by the “WHO” (World Health Organization) one-third of brain tumours are cancerous [1-4]. As stated by the “UNI” (United News of India) brain tumours are the tenth most prevalent tumor in India. As per doctors’ point of view, 90% of brain tumor cases can cure and may save many lives when detecting the brain tumor at an early stage. Hence early detection of brain tumors can increase the chance of survival [2-4].

The manual analysis of MRI reports of human brain are utilized in finding the exact boundaries of the tumors by physicians which is an intricate and demanding task Not only because of little brightness and contrast of MRI reports but also similarities of intensities among brain organelles [6]. The physical evaluation tissues of brain membrane requires ample knowledge and time-overwhelming tasks to diagnose the patient. As per the “MOHFW” (Ministry of health and Family Welfare-Government of India), there is no specific reason for brain tumors, and the possible survival rate is less than 3%. All these issues motivated the author to investigate the different automated brain tumor segmentation techniques and suitable methods to reduce the mortality rate [4-6].

In this survey, the researchers have deliberated recent and popular state of the art Deep-Learning techniques not only to segregate the variants of tumours that affect brain but also to categorize the kinds of brain tumours including Machine Learning Techniques (MLT), Artificial Intelligence (AI), and Deep-learning techniques.

Among various modalities (CT, US, and PET), the MRI modality supports the identification of brain tumors by radiologists due to low ionization and noise. To soft brain tissues MRI is more appropriate and envisages the anatomy of the brain in three different planes such as axial, sagittal, axial, and coronal view [6, 7]. Fig. 1 shows the axial, sagittal, and coronal views of the brain captured using the MRI modality or imaging technique.



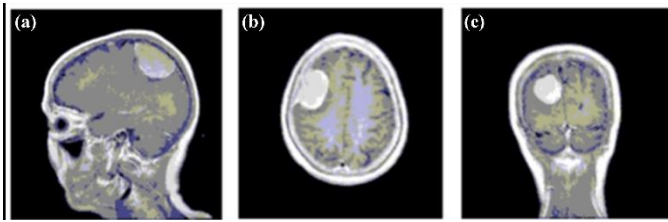


Fig. 1. Brain MRI Slices Captured from Different Directions (a) Sagittal, (b) Axial, and (c) Coronal.

MRI has various benefits over other imaging techniques, including i) high spatial resolution, ii) Functional brain measurement, iii) MRI test is acceptable for patients of any age, iv) No harmful effects on the body (no risks) due to no ionizing radiation effect, and iv) It can take images in any plan and capture finer details of soft tissues [8].

Today most of the research workout on machine learning algorithms intended to segregate the variants of tumors that affect brain automatically having capability, accuracy, reproducibility, scalability, and ease of a quantitative estimation of brain tumors [7, 9].

The methods are classified as Deep learning and Machine Learning methods. Traditional MLT use statistical learning approaches and better classify the features of low-level brain tumor. These learning methods concentrated on the estimation and localization of boundaries of the tumor [10]. The deep learning methods are requiring the smallest pre-processing steps and are more suitable for training large datasets as compared with traditional methods. Recently, in the medical field, convolutional neural networks (CNN) are more dominated than other techniques [11, 12].

As per the investigation, for automatic brain tumor detection, deep learning is a promising approach, and complex features are learned directly from input data. Deep learning approaches had been popular in the domain of computer vision due to their outstanding performance [13]. But it is required to train the samples without over-fitting and reduce the consuming time to the annotation of 3D ground truth MRI images [14].

In this survey, we have studied recent and popular Deep-Learning techniques to segment and categorize tumors from images of MRI including AI, ML methods, and DL methods.

This put forth, paper is structured accordingly: Section II is providing an overview of brain-tumor segmentation and classification; Section III and IV will provide the overview of various deep learning algorithms.

## II. TECHNIQUES FOR BRAIN-TUMOR SEGMENTATION

The process of partitioning the image (2D function) into disjoints objects and used to identify or locate the object boundaries. The ultimate goal of analysis is to detect the ROI (“region of interest”) such as location and its extension. In brain tumor segmentation techniques, the abnormal tissues are separated and identified from normal tissues [15].

Medical imaging technology plays a crucial role present in the medical field. So, the segmentation of the brain tumor region with the help of MRI scanning reports is difficult at the

primary level due to overlapping of tissues, boundary inefficiencies, dimensional differences i.e., size and shape, abnormalities, position, or location of the tumour. [16]

**Structural Segmentation:** These image segmentation techniques are depending upon the data of the structure of the desired part of the image that comes under structural segmentation techniques.

**Stochastic Segmentation:** This type of segmentation technique works on the discrete pixel value of the input image unlike structural segmentation techniques [17].

**Hybrid Segmentation Techniques:** The combination of both structural and stochastic segmentation techniques is referred to as hybrid segmentation techniques.

Depending on the human interaction, the techniques of segmentation are classified into i) manual, ii) semi-automatic and iii) automatic segmentation techniques.

### A. Manual Segmentation

In this, identifying the tumor part by the professional expert and they use a specialized tool for tumor assessment. The expert must have proper training, experience, and knowledge in the anatomy of the brain. The manual analysis of MRI reports of human brain for finding the exact boundaries of the tumors by physicians is a complicated, challenging task, and prone to error because of little or no ample brightness the reports obtained pose small amount of contrast MRI images and similarities of intensities among brain cell organelles [18]. The physical assessment of tissues in brain requires more prior knowledge and is time taking tasks to diagnose the patient. Fig. 2 depicted the edema (red-swelling), necrotic (yellow-dead), and active tumor (purple).

### B. Semi-Automatic Segmentation

This segmentation requires both human operators and computers. To initialize the segmentation process, human interaction must be required and results depend on the human operator. It consumes less time as compared with manual segmentation. Example of semi-automatic techniques is region-growing, tumor-cut method, and active contour models, etc. [18, 19].



Fig. 2. Anatomical Segmentation Manually by the Intersection.

### C. Automatic Segmentation

In this method, the human operator is not required. To solve the problem of the segmentation task, it combines both prior knowledge and artificial intelligence. There are two types of techniques as discriminating and generative methods [18,

19]. Supervised learning is one of the examples of the discriminating methods. In this, the relationship between the annotation and input image is learned through a large dataset. Unlike supervised learning, the unsupervised learning process uses the data without labels, and they are trained using the loss functions to obtain the patterns considering principal component analysis (PCA) and clustering methods [20]. Due to the complexity of (Because of high complexity (medical datasets), the machine learning techniques are not able to train the data. Recently, deep learning methods have earned a reputation due to their outstanding performance in particularly brain tumor segmentation and learning the parameters directly from data sets. The re-generative methods are utilizing the previous information involving the presence of different tumor forms [21].

### III. OVERVIEW OF DEEP LEARNING METHODS

The class of machine learning is deep learning; it can use multi-layers to learn multiple features directly from original data. In this section, the DL- concepts, techniques, and architectures for medical image analysis have been surveyed.

#### A. Neural Networks

Neural networks (NN) are the basis for deep learning methods and one type of learning algorithm. This NN has learned useful features from raw data and formed by connecting the neurons by directed links [22]. In the layers, the neurons are organized.

There are three layers (three different layers) such as i) input layer, ii) hidden layer, and iii) output layer. Typical FFNN by composing of three layers shown in Fig. 3.

In input layer each neuron is connected to another neuron in the upstream layer's output. Similarly, each neuron's output is coupled to all the neurons in the downstream layer's input. The weight is adjusted in each link during the learning process [22, 23]. The network's topology immediately forms an acyclic graph, and the network is known as a "FFNN" which refers to Feed Forward Neural Network" in which every neuron is linked or interconnected to the neurons in next layer. The deep neural network is made up of multiple layers, or hidden layers.

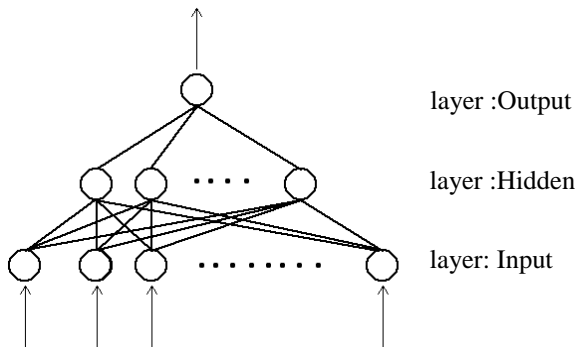


Fig. 3. Typical Structure of FFNN with Three Layers.

Recently, in a supervised manner, all the methods are trained to make easier the training procedure. There are more fashionable architectures utilized in the analysis of health care domain: convolution neural networks (CNNs) and recurrent neural networks (RNNs) [21, 24]. The CNNs are gaining

massive popularity to solve problems in medical field as compared to the RNNs. The overview of these methods is given in the following section.

#### B. Convolutional Neural Networks (CNNs)

The CNNs are used to work on convolutional operations and one type of neural network. There are two types of methods such as traditional and DL methods. The learning approaches in statistics are applied for the classification of low-level brain tumors, which is a regular task in conventional machine learning methods. The CNNs are the dominant area in the last years for segmentation of brain tumors and requires fewer pre-processing techniques and is suitable for training the large datasets than conventional techniques [25].

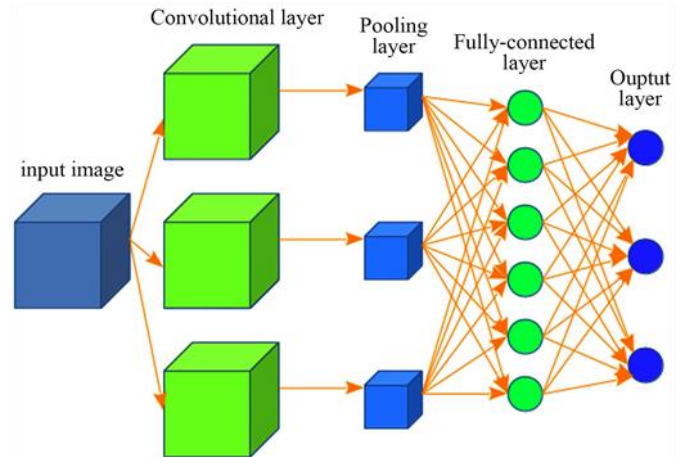


Fig. 4. Typical Structure of CNN.

Nowadays, to solve computer and medical image problems CNN is one among the dominant models, especially for segregation of diverse kinds of brain tumours. The CNN has many layers as shown in Fig. 4 that are transforming the input images into output (normal/abnormal) by using convolutional filters while learning the high-level features [26]. The CNN models have been learning the spatial features in the given data. The first convolutional layer can learn the low edges and second layer will learn high-level features. Next, the units of convolutional layers shrink the number of parameters to learn by distributing the weights. Thereby, increase the efficacy of the network [27]. Node graph for CNN is indicated in Fig.5 (a)

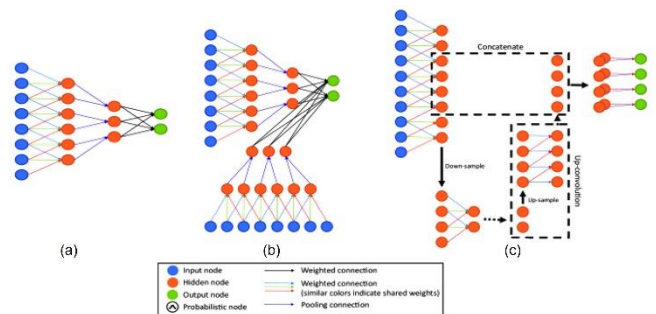


Fig. 5. Node Graph of 1D Representation of CNN Architecture used in Medical Imaging. (a) CNN, (b) Cascaded CNN, and (c) U-net.

The MLTs and CNNs majorly differ where the network shares the weights to perform the convolutional operation, no

need for separate detectors to learn, and weights do not depend on the original image size [28]. And also, due to the pooling layer in the CNNs the neighborhood pixel values are accumulated by using the “max or mean” functions [27, 28]. Regular neural networks (completely connected layers) are attached at the conclusion of CNN, but the weights are not shared. The distribution over classes is constructed by applying the SoftMax function and sending the activations into the final layer. Maximum likelihood [29] is used to train the network.

The layers of convolution are learning the local and complex features in the hierarchy from the original given data. To summarize the key features, the Pooling-Layer is included in the middle of consecutive convolutional layers to decrease the enormous parameters and then forwarded to downstream layers. The translation invariant was created to identify the learned patterns, irrespective of geometric transformations [30].

#### IV. DCNN: DEEP CONVOLUTION NEURAL NETWORK

There are different architectures proposed by researchers such as single pathway, dual pathway, cascaded, and U-net. All these architectures are briefed below:

##### A. Single Pathway

The architecture of a single pathway is looking like a feed-forward deep neural network (FFDNN) and it is a basic network for remaining architectures. In this single path, information is passed down from one of the input layers to the other layer of classification. The 3-D ‘single path CNN’ was proposed by Urban et al. [31]. This architecture consists of the completely connected convolutional layer as a classification layer and can classify multiple 3-D pixels into one. In [32] every image from the sensory system is supplied to the different 2-D CNN and features from the results of CNN are utilized to train a random forest classifier. The neighbourhood information is obtained from XX, YZ, and XZ planes near each center pixel. In convolutional layers, small kernels have been used by Pereira et al. [33], which can learn more features by obtaining the very deep and deep medic networks. The proposed architectures got a 1st and 2nd place in ‘BraTs-2013’ and ‘BraTs-2015’ challenges respectively.

##### B. Dual Pathway

Pixel-wise classification will be performed in many segmentation approaches, here extract the input patches from input MRI image, and then without considering the neighborhood information central pixels labels were predicted. Infiltrating the process can be risky and makes ambiguous boundaries. So, to achieve better results only local information is not sufficient. To avoid this problem, the authors [34] mix the neighborhood information by employing CNN with dual-path data streams. These dual paths were mixed to impact each pixel label prediction. Among two paths. The visual elements of the region near each center pixel are represented by one of the pathways. The second stream will relate to global information, and it will include the location of the discovered patch in the brain [35].

##### C. Cascaded Architecture

This architecture will make the multi-scale label prediction independently from others as compared with the dual-path way. The output of the CNN was chained with the other. There are several architectures, among them, input cascade is one of the most important architectures and used to chain the secondary CNN with contextual information. The typical multi-stream CNN is shown in Fig. 5 (b). The local pathway concatenation is a cascaded architecture in which the first CNN’s output is sequenced and added with second CNN’s first hidden layer output regardless of its input [36].

Another important cascaded architecture is the hierarchical segmentation [37], in which brain tumor segmentation of brain tumor region was accomplished by decreasing the “multi-class segmentation challenge” into a “multi-stage binary segmentation” problem. It uses the hierarchical architecture of tumor sub-regions to reduce false positives while simultaneously resolving the inherent imbalance problem. The entire tumor was segmented from the reports of MRI which has been given as inputs at the initial stage of the design, and then a boundary box was used in the second step. Next, separate the remaining sub-regions using either multi-class intra-tumor segmentation or successive binary segmentation. [38].

##### D. U-net

The U-net architecture [39] is constructed exclusively for biomedical image segmentation and looks like an encoder and decoder network. A U-shaped design consisting of an encoder at the contracting path and decoder at the expanding side which entirely builds up the U-net. In the contracting pathway, the ReLU layer and the max-pooling layer come after the two convolutional layers. The spatial data decreases as the path contracts, while the feature information increases. From contracting to skip connecting, the expanding path comprises a sequence of up-sampling processes paired with high-resolution features. The typical U-net showed in Fig. 5 (c).

#### V. CNN MODELS FOR BRATS

The current methods for segmenting and classifying kinds of tumours that affect human brain from MRI data is discussed in this section. The manual assessment tissue organelles of brain require an ample prior knowledge, are time taking tasks, and prone to error during the diagnosis process. Therefore, these issues motivated to development of automated tumor detection techniques by several researchers, and it becomes a significant area for research in medical image processing. This review article mostly concentrates on the segmentation and classification processes based on traditional AI, MLT, and DLM in the last five years.

The literature has highlighted several automated approaches in the domain of health care image analysis to diagnose the health issues such as tumours, lung cancer [19], skin cancer [20, 21], and more [22, 23]. Many strategies for pre-recognition and categorization of brain tumors are offered as a result of all of them.

Kumar Agrawal, Ullas, and Pankaj Kumar Mishra [40] studied various state-of-the-art algorithms for detecting and classifying tumors accurately. They revealed that deep learning

may be used in conjunction with several transfer learning approaches to construct a systematic and efficient approach for the early identification of tumours in brain in this proposed study.

Rehman, Amjad, and colleagues [41] suggested a novel DL-based technique for classifying microscopic brain tumours. The authors designed the 3D architecture of CNN, draw out the brain tumour and then send it to the pre-trained CNN model for parameter extraction. Next, to choose best features a correlation-based selection approach is employed. The chosen parameters are validated by employing the FFNN for the last classification. The authors utilized BraTS 2015, 2017, and 2018 for validation and achieved more than 92.67% of accuracy.

Sharif, Muhammad Imran, et al. [42] presented a new automated DL- method to classify the multi-class brain tumors. In the proposed method, the densenet201 pre-trained DL model was trained by deep transfer of imbalanced data learning. The average pool layer is used to retrieve the training model features. Two methods are used to pick the features. i) entropy-kurtosis-based high feature values (EKbHFV) and ii) metaheuristic-based modified genetic algorithm (MGA). The non-redundant serial-based approach is used to fuse the EKbHFV and MGA-based features. Finally, a multiclass SVM cubic classifier is used. They concluded that the presented method has achieved an accuracy of about 95.5%.

Khan, Amjad Rehman, et al. [43] presented new DL techniques to classify the tumours from MRI brain images. The author's method consists of pre-processing, segmentation using the K-means technique, and classification using the fine-tuned nineteen-layered visual geometric group (VGG19) model. To enhance the scale of the available data, the synthetic data augmentation idea is presented. The Put forth method outperformed earlier when compared with accuracy, the put forth model outperformed the previous state-of-the-art approach.

Khairandish, Mohammad Omid, and colleagues [44] created a hybrid model that uses CNN and SVM for classification and threshold-based segmentation for detection. To categorize benign and malignant tumors, the authors used a publicly available dataset. The suggested hybrid CNN-SVM technique achieves an overall accuracy of 98.4959 percent.

TAS, Muhammed Oguz, and Semih ERGİN [45] in this survey, the authors studied the segmentation of tumors from abnormal brain MRI images with DL and K-means approach. The proposed method was extracting the tumor area automatically with an accuracy and sensitivity of 84.45% and 95.04% respectively.

The authors [46] have been proposed the Google Net approach depending upon the CNN DL approach to classify the different types of tumors from MRI brain images and to overcome the difficulties in the classification of attributes such as variants in texture, size, and shape. The authors use MRI datasets to perform five-fold cross-validation and authenticated the put forth system's performance in terms of "area under the curve" (AUC), F-score, recall, precision, and specificity. The proposed system with transfer learning provides the

classification accuracy of 97.8% and 98% with the multiclass SVM method.

In [47], the stationary wavelet transform (SWT) approach and modern growing convolution neural network (GCNN) (alarmed CNN) was developed by the authors in this study to improve the efficiency of an automated brain tumor segmentation system. SVM and CNN have done better than the suggested work in all aspects.

For automatic segmentation, the authors [48] offered enhanced convolutional neural networks (ECNN) with loss function optimization via the BAT algorithm. They presented the optimization-based MRIs image segmentation. To overcome the overfitting problem, assigned the lesser weights to the network. The efficacy of the proposed system was authenticated by utilizing the different popular brain tumor datasets. The overall results indicate that the presented method shows better performance.

The segmentation in a DL approach [49] is done with CNN. The Convolution Neural Network has deep architecture; this approach employs three tiny kernels. For the pre-processing of pictures, intensity normalization and data augmentation were used. The method is evaluated using the famous dataset (BraTS 2013 and BraTS 2015).

The authors [50] presented a solution for the low accuracy of brain tumor segmentation using DL techniques. They studied MRI images in different angles and applied different networks for segmentation. Evaluated the effect of separate networks and compare them with a single network. The dice score 0.73 and 0.79 is achieved with single and multiple networks, respectively.

The DCNN-F-SVM was presented by Wu Wentao, et al. [51] for segregation of various types of brain tumour. The proposed segmentation model has three stages: first, DCNN is trained, then predicted labels are produced from the trained DCNN, and finally, the DCNN and an integrated SVM is a deep classifier which is connected in series. They used the BraTS and self-made datasets to run each model for brain tumors segmentation. The authors conclude the presented methodology has performed for brain tumor detection better than DCNN and SVM classifier.

Sun, Li, et al. [52] presented segmentation of brain tumor and glioma survival anticipation utilizing the based framework. For tumor segmentation, the researchers used multimodal MRI scans and three 3D CNN designs. For survival prediction, 4,524 radiomic characteristics are retrieved from segmented regions, and potent features are opted making use of both decision tree and cross-validation techniques. They trained a random forest model to predict the patient's survival rate. The authors were preferred BraTS 2018 and achieved 61.0% of classification accuracy for short, mid, and long survivors among 60+ participated teams.

Bhandari, Abhishta, Jarrad Koppen, and Marc Agzarian [53] investigated the potential use of CNNs by studying radiomics and analyzed quantifiable features tumours including texture, shape and ability to forecast clinical consequences such as survival and diagnosis. The authors also investigated the role of CNNs intended for brain tumors segmentation



through the education viewpoint and performed the literature review.

Sharif, Muhammad Irfan, et al. [54] presented a non-passive learning feature selection strategy for detecting and distinguishing brain cancers. The contrast enhancement is first fed into SbDL for the development of the saliency map, and then it is transformed to binary using thresholding algorithms. Deep feature extraction was done using the inception V3 pre-trained CNN model was utilized for extraction of deep features for the purpose of classification. Next, for better texture analysis, the features are sequenced and aligned with dominated rotated (DRLBP) and particle swarm optimization (PSO) is used to optimize the concatenated vector. The SbDL segmentation and classification strategy are applied on BARTS 2017 and BRATS 2018 to validate. The presented method outperforms for classifying and segregation of brain tumours) with accuracy of 93.7% and 92% respectively.

Khan, Muhammad A., et al. [55] developed an automated system based on marker-based watershed segmentation for extraction and classification of brain tumors using MRI images. The first contrast of the tumor was enhanced by using the gamma contrast stretching technique and then the segmentation process was performed using the marker-based watershed algorithm to detect the tumor exactly. Next, by using the chi-square max conditional priority feature method, the features are selected and then fused by the serial-based concatenation method before classification. The SVM was applied to classify the tumors using the datasets such as Harvard, BRATS-2013, privately collected. The overall results revealed that the presented system performs than existing systems with high accuracy.

## VI. DATASETS

On selected datasets of tumor MRI scans, all of the strategies described in this work were tried. The research groups interested in automatic tumor segmentation from abnormal brain images over the last five years. The researchers have used the different private and public datasets to evaluate the various algorithms. A number of datasets are accessible for training and testing purposes. The challenges of benchmark datasets provide the publicly available datasets such as DICOM [56], BRATS [57], BRATS 2013, 2013, 2015, 2016, 2017, 2018, and 2020), MICCAI [58], Brain Web [59], Harvard business school [60], Internet Brain Segmentation Repository (IBSR)[61],nyrosynth.org [62], ABIDE [63], National Bioscience Database Center (NBDC) [64], Med Pix, PGIMER dataset [65], SPL database [66] etc.

## VII. PERFORMANCE EVALUATION METRICS

The effectiveness of segmentation or classification methods can be measured in a number of ways. To demonstrate their results, the authors employ various performance measure parameters. The analysis of traditional methods is commonly evaluated by mean square error (MSE), peak signal to noise ratio (PSNR), entropy, and correlation. Among various image quality assessment parameters for analysis of the result, some of the overlapped based parameters are briefed below:

**Accuracy:** The ability to determine the precision or proximity of the tumor is referred to as accuracy. The following factors influence it:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (1)$$

True positive is denoted by TP, whereas true negative is denoted by TN. False-positive and false-negative are represented by the letters FP and FN, respectively [56].

**Precision:** It denotes the consistency of two or more values. The precision formula is as follows:

$$precision = \frac{TP}{TP + FP} \quad (2)$$

Where TP represents the true positive and FP represents the false positive. The fraction of true positives is referred to as precision. The valid positive results are evaluated by dividing valid positive outcomes aided by the segmentation algorithm. Similarly, the pixels are break down into the cluster and pixels that are of that cluster

Sensitivity or recall or true positive rate: It's the ability to find a tumor or any other affected location [50]. The mathematical equation is written as follows:

$$sensitivity = \frac{TP}{TP + FN} \quad (3)$$

The TP denotes true positive and FN denotes false negative. Sensitivity is the proportion of correctly segmented images to all segmented images. The greatest results for accuracy, precision, and sensitivity suggest that the brain tumor can be recognized precisely and without ambiguity.

**Confusion Matrix:** It is used to give required information about the actual and predicted results by a particular method. The confusion matrix revealed in Table I that was seen below:

TABLE I. THE REPRESENTATION OF THE CONFUSION MATRIX

Type	Predicted class 1	Predicted class 2
Actual class 1	T <sub>P</sub>	F <sub>N</sub>
Actual class 2	F <sub>P</sub>	T <sub>N</sub>

**SDE-Segmentation Distance Error:** It is used to assess the effectiveness of segmentation procedures and the equation is denoted as

$$SDE = \frac{\|\varphi F - \varphi D\|^2}{\|\varphi D\|^2} \quad (4)$$

Where  $\varphi F$  is the terminal contour and  $\varphi D$  is the aspired contour derived from the brain tumor ground truth image. The SDE returns the normalized contour between the intended and terminal contours. The SDE range from 0 to 1, here 1 indicates the inadequate segmentation [32].

**Jaccard Similarity Index (JSI):** The JSI is defined as the ratio of common voxels in the input image (X) to the union function, or the collection of voxels in the input image (X) and segmented output image (Y) [8, 17]. The JSI mathematical equation is given by:

$$JSI = J(X, Y) = \frac{|X \cap Y|}{|X \cup Y|} \quad (5)$$

It is a scale that runs from 0% to 100% in terms of similarity between the input image and the segmented image. The greater the similarity, the higher the percentage.

**SSIM -Structural Similarity Index Measure:** The SSIM is a perceptual parameter, which means that image quality may suffer as a result of data compression, lack of data transport, or other image processing procedures. This expression is provided by:

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \quad (6)$$

Where  $\mu_x$  and  $\mu_y$  are the mean,  $\sigma_x$  and  $\sigma_y$  are the variance, and  $\sigma_{xy}$  is the covariance of x and y. The values  $c_1$  and  $c_2$  are constants. A larger SSIM value guarantees improved brightness, contrast, and structural material quality [55]. In addition to these characteristics, computation time and iteration count are utilized to calculate the performance of the suggested approaches.

### VIII. HARDWARE AND SOFTWARE TOOLS

Nowadays various open-source softwares are used by researchers to speed-up the deep learning systems. This section has covered a brief description of hardware and software used in research papers.

For deep learning purposes, there is the availability of computing libraries such as GPU and CPU. The GPUs have been performed parallel computation with a high execution rate as compared with the CPUs. The GPUs hardware in deep learning is 10-30 times faster than CPUs. These libraries of GPUs also provide various operations implementation in NN i.e., convolutions and user friendly. Due to the popularity of DL, there is more availability of open-source software packages.

Caffe [67] is first established for computer vision applications by graduate students (Jia et al., 2014) at Berkeley and supports C++ and Python interfaces. This deep learning framework is not only used in computer vision applications, but it can also use in other fields such as robotics, neuroscience, and astronomy. For deep learning, from training to architecture development, provides the complete tool kit with good examples. It allows the user to implement the building models and models of deep learning with various algorithms.

TensorFlow [68], developed by Google (Abadi et al., 2016) for large-scale machine learning applications and supports the

C++ and interfaces of Python. It is end-to-end distributed deep learning and supports the data flow graphs execution in mobile devices or heterogeneous devices. It is designed for fast experimentation with a deep learning model using a complete toolbox and simply the parallelism of the model.

Theano [69] was built in the Montreal lab called MILA (Bastien et al., 2012) and offers Python interfaces. It is used to execute and compile the mathematical expressions by syntax NumPy quickly using both GPUs and CPUs, especially for large-scale dataflow. The other high-level software packages constructed upon the Theano consider Pylearn2, Keras, and Lasagne.

Pytorch [70] is an important approach to construct the computational graph dynamically rather than static computational graph before running the model and open-source framework of deep learning. It is flexible, powerful, and easy of debugging (Collobert et al., 2011). The Pytorch is suitable for the production and execution of models on edge devices. It is used in Facebook AI research.

Pylearn2 [71] allows the user to construct or implement the machine learning models in an arbitrarily and free (open source) machine learning library. This library is flexible and easy to use. But unfortunately, due to the lack of active developers, it is fallen as compared with other frameworks.

Keras [72] is one of the rapidly developing application programming interfaces (API) for various applications of deep learning and supports multiple data-flow graphs like Theano. To run the experiments with models Keras consist of simple APIs which has provision to run in mobile devices and also in browsers. This platform was adopted for research areas and industrial applications due to its simplicity of usage (user-centric approach).

Lasagne [73] In Theano, the Lasagne is a trivial library for building and training NNs. To run the Lasagne, you will need Python 2.7 or 3.7, and instructions are running in MAC or Linux systems. It has the capability of powerful mathematical computations and constructed upon the Theano.

### IX. SUMMARY AND DISCUSSION

Deep Learning (DL) based approaches for segregation brain tumour have recently sparked a lot of interest. In brain tumour segmentation, deep learning systems are trained on big datasets to segment the tumour from MRI images by learning a hierarchy of complicated properties straight from data. As a result, CNN-based models are the most used in medical image analysis, with success in fields including natural language processing, audio identification, and brain tumour segmentation. For instance, in Fig. 6, CNN was used to segment brain tumors using a single node into CNN. Input, convolution with nonlinearity correction using ReLU, overfitting correction using pooling, feature map flattening into a column, and finally insertion into the neural network.



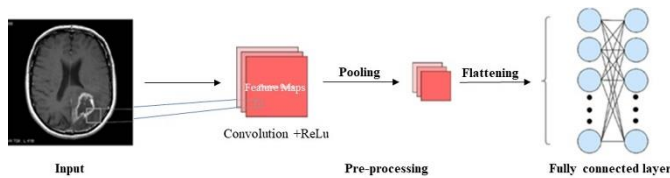


Fig. 6. Input the Image into a Single Node within a CNN.

In this work, recently published papers have been reviewed. It is clear that DL methods play a crucial role to address the

problems faced in automated brain tumor segmentation. Pre-trained CNNs are employed as feature extractors in recent studies, and these networks are installed and employed directly on the medical images for a specific purpose. In the previous two years, end-to-end trained CNNs have been favored for medical image analysis interpretation. Nowadays, deep learning approaches are integrated with traditional machine learning handcrafted methods. In recent works uses the U-net and ensemble methods to resolve the problem of segmentation of a brain tumor [74-78].

TABLE II. THE COMPARISON OF VARIOUS DEEP LEARNING TECHNIQUES BASED ON DATASETS AND ACCURACY

Author, Year	Method	Dataset	Accuracy (%)
K. Agrawal, 2021[40]	Transfer Learning method	BraTs 2018	97.17
Rehman, 2021[41]	3D-CNN Architecture	BraTs 2018	95.53
Sharif, 2021 [42]	Entropy–Kurtosis-based High Feature Values (EKbHFV) and modified genetic algorithm (MGA)	BraTs 2019	95.0
Khan, 2021 [43]	Deep learning approach- finetuned VGG19	BraTS 2015	94.06
Khairandish, 2021 [44]	hybrid CNN-SVM	BraTs 2019	98.49
TAS, 2020 [45]	Traditional Deep Learning Technique		84.45
Wu, 2020 [51]	DCNN-F-SVM	BraTs 2018	96.0
Sharif, 2020 [54]	Pixel Increase along with Limit (PlaL)	BraTs 2018	93.7
S. Deepak, 2019 [46]	deep CNN features via transfer learning	BraTS 2017	98
Mamta Mittal, 2019 [47]	Growing Deep Convolutional Network (GCNN)	BraTs 2018	97.7
Thaha, 2019 [48]	Enhanced Convolutional Neural Networks (ECNN)	BraTs 2015	92.0
Sun Li, 2019 [52]	3D CNN architectures (Cascaded Anisotropic CNN, German Cancer Research Center NET - DFKZ Net, 3D-U-Net)	BraTs 2018	91.0
Sobhaninia, 2018 [50]	LinkNet network	BraTs 2015	89.12

The ensemble approaches improve the robustness of every approach by combining the results of segmentation and providing better performance as compared to several models. The single U-net-based models support the argument. The overview of recent approaches used for brain tumor segmentation along with its accuracy is shown in Table II. For deep learning algorithms, we need an enormous quantity of training data to simplify properly invisible data and poses many challenges in the domain of medicine. It requires an experienced neuro-radiologist before applying to the supervised training.

So, it's an expansive, large memory resources, and time-consuming task. But recently the BraTS challenges provides training and testing to users and due to proper training, the over-fitting problem will be reduced. At the same time, the researchers have implemented data augmentation to avoid the problem of unavailability of large-scale datasets. The computational and memory requirements increased further due to 3D deep learning models.

As per the literature, the authors have used costly mathematical functions, well software libraries, multi-GPU environments and train the data in a distributed manner. To enhance the correctness durability of segmentation algorithms, authors must carefully initialize the hyper-parameters, employ proper pre-processed approaches and use Latest training methods

## X. CONCLUSION

We discussed methods for segmenting brain tumors, different architectures of deep learning methods for automatic

brain tumor segmentation, a literature review of recently published papers, tools for implementing the algorithms, dataset availability, and appropriate performance metrics to estimate the performance of each method in this paper. As compared with traditional techniques, the deep learning methods are still superior due their robustness and performance. The novel architectures of Deep-Learning have a great potential to avoid the inherent class imbalance problems in tumor segmentation by using proper pre-processing, initialization of weights, and sophisticated training methods. In many segmentation techniques, due to the lack of a large-scale training dataset, its performance will be degraded.

This paper contains an overview of contemporary strategies for segmenting and classifying brain tumors using MRI data. The goal of the presented survey is to demonstrate briefly about the most commonly used strategies for segmenting and classifying tumours. Various tumor segmentation techniques (manual, semi-automatic and automatic), deep learning methods (NN, CNNs, and DCNN), and different architectures are used in DCCN.

This paper consists of a review of the recently published research articles from science direct, IEEE explore, etc. The literature covered the ML techniques, DL methods, and hybrid methods for abnormality segmentation from MRI brain images. This paper also contains publicly available datasets or benchmarking challenges such as BraTS 2012-202, MICCAI, Harvard University, and Brain Web. As per the literature of various published papers, the authors have been primarily used datasets are BraTs-2013 followed by BraTs-2015.

REFERENCES

- [1] Amin J, Sharif M, Yasmin M, Fernandes SL. A distinctive approach in brain tumor detection and classification using MRI. *Pattern Recognition Letters*. (2017) pp. 1-10.
- [2] Tiwari, Arti, Shilpa Srivastava, and Millie Pant. "Brain tumor segmentation and classification from magnetic resonance images: Review of selected methods from 2014 to 2019." *Pattern Recognition Letters* 131 (2020): 244-260.
- [3] Magadza, Tirivangani, and Serestina Viriri. "Deep Learning for Brain Tumor Segmentation: A Survey of State-of-the-Art." *Journal of Imaging* 7.2 (2021): 19.
- [4] Litjens, Geert, et al. "A survey on deep learning in medical image analysis." *Medical image analysis* 42 (2017): 60-88.
- [5] U. Sai Deepthi, A. Sudha Madhuri, P. Sai Prasad., "Comparative Analysis of Brain Tumour Detection Using Deep Learning Methods", 2019, *International Journal of Scientific & Technology Research*, vol.8, issue.12, pp:250-254.
- [6] Akram, S.U., Kannala, J., Eklund, L., Heikkilä, J., 2016. Cell segmentation proposal network for microscopy image analysis. In: *Proceedings of the Deep Learning in Medical Image Analysis (DLMIA)*. In: *Lecture Notes in Computer Science*, 10 0 08, pp. 21–29. doi: 10.1007/978-3-319-46976-8\_3.
- [7] P. V. Nagajaneyulu, and K. Satya Prasad, "Brain Tumor Segmentation of T1w MRI Images Based on Clustering Using Dimensionality Reduction Random Projection Technique," *Current Medical Imaging*, vol.17, no.3, pp:1-11, March.2021, DOI:10.2174/1573405616666200712180521
- [8] S. Deepak, P.M. Ameer, Brain tumor classification using deep CNN features via transfer learning, *Computers in Biology and Medicine*, 111, (2019).
- [9] Liu J, Li M, Wang J, Wu F, Liu T, and Pan Y. A Survey of MRI-Based Brain Tumour Segmentation Methods. *Tsinghua Science & Technology*. 2014; 19(6), 578-595.
- [10] Wadhwa, A.; Bhardwaj, A.; Singh Verma, V. A review on brain tumor segmentation of MRI images. *Magn. Reson. Imaging* 2019, 61, 247–259.
- [11] Muhammad, K.; Khan, S.; Ser, J.D.; de Albuquerque, V.H.C. Deep Learning for Multigrade Brain Tumor Classification in Smart Healthcare Systems: A Prospective Survey. *IEEE Trans. Neural Netw. Learn. Syst.* 2020, 1–16.]
- [12] Alom, M.Z.; Taha, T.M.; Yakopcic, C.; Westberg, S.; Sidike, P.; Nasrin, M.S.; Hasan, M.; Van Essen, B.C.; Awwal, A.A.S.; Asari, V.K. A State-of-the-Art Survey on Deep Learning Theory and architectures. *Electronics* 2019, 8, 292
- [13] Zikic, D.; Ioannou, Y.; Brown, M.; Criminisi, A. Segmentation of Brain Tumor Tissues with Convolutional Neural Networks. In *Proceedings of the BRATS-MICCAI*, Boston, MA, USA, 14 September 2014; pp. 36–39.
- [14] Urban, G.; Bendszus, M.; Hamprecht, F.A.; Kleesiek, J. Multi-Modal Brain Tumor Segmentation Using Deep Convolutional Neural Networks. In *Proceedings of the BRATS-MICCAI*, Boston, MA, USA, 14 September 2014; pp. 31–35
- [15] Havaei, M.; Guizard, N.; Larochelle, H.; Jodoin, P.M. Deep Learning Trends for Focal Brain Pathology Segmentation in MRI. In *Machine Learning for Health Informatics*; Holzinger, A., Ed.; Springer International Publishing: Berlin/Heidelberg, Germany, 2016; Volume 9605, pp. 125–148
- [16] P.V. RohiniIana and M. Pushpa ani. Analysis and Detection of Brain Tumour Using Image. Processing Technique. *International Journal of Advanced Technology in Engineering and Science*. 2015; 3(1), 393-399.
- [17] Rajesh Babu, K., P. V. Nagajaneyulu, and K. Satya Prasad, "Performance Analysis of CNN Fusion Based Brain Tumour Detection Using Active Contour Segmentation Techniques," *International Journal of Signal and Imaging Systems Engineering*, vol. 12, no.1, pp:62-70, March 2020. DOI: 10.1504/IJSISE.2020.113571
- [18] Isn, A.; Direko glu, C.; Sah, M. Review of MRI-Based Brain Tumor Image Segmentation Using Deep Learning Methods. *Procedia Comput. Sci.* 2016, 102, 317–324.]
- [19] Saleha Masood et al. A Survey on Medical Image Segmentation. *Curr. Med. Imaging Rev.* 2015; 11(1), 3-14.
- [20] Saleha Masood, Muhammad Sharif, Afifa Masood, Mussarat Yasmin, and Mudassar. A Survey on Medical Image Segmentation. *Curr. Med. Imaging Rev.* 2015; 11(1), 3-14.
- [21] Isn, A.; Direko glu, C.; Sah, M. Review of MRI-Based Brain Tumor Image Segmentation Using Deep Learning Methods. *Procedia Comput. Sci.* 2016, 102, 317–324.] [Svozil, D.; Kvasnicka, V.; Pospichal, J. *Introduction to Multi-Layer Feed-Forward Neural Networks*. Chemom. Intell. Lab. Syst. 1997, 39, 43–62
- [22] Chen, L.; Bentley, P.; Mori, K.; Misawa, K.; Fujiwara, M.; Rueckert, D. DRINet for Medical Image Segmentation. *IEEE Trans. Med. Imaging* 2018, 37, 2453–2462.
- [23] Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning; Adaptive Computation and Machine Learning; The MIT Press: Cambridge, MA, USA*, 2016.
- [24] Alom, M.Z.; Taha, T.M.; Yakopcic, C.; Westberg, S.; Sidike, P.; Nasrin, M.S.; Hasan, M.; Van Essen, B.C.; Awwal, A.A.S.; Asari, V.K.
- [25] Alom, M.Z.; Taha, T.M.; Yakopcic, C.; Westberg, S.; Sidike, P.; Nasrin, M.S.; Hasan, M.; Van Essen, B.C.; Awwal, A.A.S.; Asari, V.K. A State-of-the-Art Survey on Deep Learning Theory and architectures. *Electronics* 2019, 8, 292.
- [26] A State-of-the-Art Survey on Deep Learning Theory and architectures. *Electronics* 2019, 8, 292.
- [27] Cheng, J.-Z., Ni, D., Chou, Y.-H., Qin, J., Tiu, C.-M., Chang, Y.-C., Huang, C.-S., Shen, D., Chen, C.-M., 2016a. Computer-aided diagnosis with deep learning architecture: applications to breast lesions in US images and pulmonary nodules in CT scans. *Nat. Sci. Rep.* 6, 24454. doi: 10.1038/srep24454
- [28] Chollet, F. *Deep Learning with Python; Manning Publications Co.: Shelter Island, NY, USA*, 2018.
- [29] Muhammed Talo, Ulas Baran Baloglu, ÖzalYıldırım, U Rajendra Acharya, Application of deep transfer learning for automated brain abnormality classification using MR images, *Cognitive Systems Research*, 54, (2019), pp. 176-188
- [30] Tianbao Ren, Huanhuan Wang, Huilin Feng, Chensheng Xu, Guoshun Liu, Pan Ding, Study on the improved fuzzy clustering algorithm and its application in brain image segmentation, *Applied Soft Computing*, 81, (2019).
- [31] Urban, G.; Bendszus, M.; Hamprecht, F.A.; Kleesiek, J. Multi-Modal Brain Tumor Segmentation Using Deep Convolutional Neural Networks. In *Proceedings of the BRATS-MICCAI*, Boston, MA, USA, 14 September 2014; pp. 31–35
- [32] Rao, V.; Sarabi, M.S.; Jaiswal, A. Brain tumor segmentation with deep learning. In *Proceedings of the MICCAI Multimodal Brain Tumor Segmentation Challenge (BraTS)*, 2015; pp. 56–59.
- [33] Pereira, S.; Pinto, A.; Alves, V.; Silva, C.A. Brain Tumor Segmentation Using Convolutional Neural Networks in MRI Images. *IEEE Trans. Med. Imaging* 2016, 35, 1240–1251.
- [34] Casamitjana, A.; Puch, S.; Aduriz, A.; Sayrol, E.; Vilaplana, V. 3D Convolutional Networks for Brain Tumor Segmentation. In *Proceedings of the MICCAI Challenge on Multimodal Brain Tumor Image Segmentation (BraTS)*, 2016; pp. 65–68. Available online: <https://imatge.upc.edu/web/sites/default/files/pub/cCasamitjana16.pdf>.
- [35] Havaei, M.; Davy, A.; Warde-Farley, D.; Biard, A.; Courville, A.; Bengio, Y.; Pal, C.; Jodoin, P.M.; Larochelle, H. Brain tumor segmentation with Deep Neural Networks. *Med. Image Anal.* 2017, 35, 18–31.
- [36] Hussain, S.; Anwar, S.M.; Majid, M. Brain Tumor Segmentation Using Cascaded Deep Convolutional Neural Network. In *Proceedings of the 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Seogwipo, Korea, 11–15 July 2017; pp. 1998–2001.
- [37] Pereira, S.; Oliveira, A.; Alves, V.; Silva, C.A. On hierarchical brain tumor segmentation in MRI using fully convolutional neural networks: A preliminary study. In *Proceedings of the 2017 IEEE 5th Portuguese Meeting on Bioengineering (ENBENG)*, Coimbra, Portugal, 16–18 February 2017; pp. 1–4.

- [38] Wang, G.; Li, W.; Ourselin, S.; Vercauteren, T. Automatic Brain Tumor Segmentation Based on Cascaded Convolutional Neural Networks with Uncertainty Estimation. *Front. Comput. Neurosci.* 2019, 13, 56.
- [39] Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. *arXiv 2015*, arXiv:1505.04597.
- [40] Kumar Agrawal, Ullas, and Pankaj Kumar Mishra. "Classification and Detection of Brain Tumor Through MRI Images Using Various Transfer Learning Techniques." *Annals of the Romanian Society for Cell Biology* 25.6 (2021): 5484-5491
- [41] Rehman, Amjad, et al. "Microscopic brain tumor detection and classification using 3D CNN and feature selection architecture." *Microscopy Research and Technique* 84.1 (2021): 133-149.
- [42] Sharif, Muhammad Imran, et al. "A decision support system for multimodal brain tumor classification using deep learning." *Complex & Intelligent Systems* (2021): 1-14.
- [43] Khan, Amjad Rehman, et al. "Brain tumor segmentation using K-means clustering and deep learning with synthetic data augmentation for classification." *Microscopy Research and Technique* (2021).
- [44] Khairandish, Mohammad Omid, et al. "A Hybrid CNN-SVM Threshold Segmentation Approach for Tumor Detection and Classification of MRI Brain Images." *IRBM* (2021).
- [45] TAS, Muhammed Oguz, and Semih ERGİN. "Detection of the Brain Tumor Existence Using a Traditional Deep Learning Technique and Determination of Exact Tumor Locations Using K-Means Segmentation in MR Images." *İleriMühendislikÇalışmalariveTeknolojileriDergisi* 1.2: 91-97.
- [46] S. Deepak, P.M. Ameer, Brain tumor classification using deep CNN features via transfer learning, *Computers in Biology and Medicine*, 111, (2019).
- [47] Mamta Mittal, Lalit Mohan Goyal, Sumit Kaur, Iqbaldeep Kaur, Amit Verma, D. Jude Hemanth, Deep learning-based enhanced tumor segmentation approach for MR brain images, *Applied Soft Computing*, 78, (2019), pp. 346-354.
- [48] Thaha, M. Mohammed, Kumar, K. Pradeep Mohan, Murugan, B. S., Dhanasekaran, S., Vijayakarthish, P., Selvi, A. Senthil, Brain Tumor Segmentation Using Convolutional Neural Networks in MRI Images, *Journal of Medical Systems*, 43 (9), (2019), pp. 1-10.
- [49] Alzubaidi, L., Zhang, J., Humaidi, A.J. et al. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. *J Big Data* 8, 53 (2021). <https://doi.org/10.1186/s40537-021-00444-8>
- [50] Sobhaninia, Zahra, et al. "Brain tumor segmentation using deep learning by type-specific sorting of images." *arXiv preprint arXiv:1809.07786* (2018).
- [51] Wu, Wentao, et al. "An intelligent diagnosis method of brain MRI tumor segmentation using deep convolutional neural network and SVM algorithm." *Computational and Mathematical Methods in Medicine* 2020.
- [52] Sun, Li, et al. "Brain tumor segmentation and survival prediction using multimodal MRI scans with deep learning." *Frontiers in neuroscience* 13 (2019): 810.
- [53] Bhandari, Abhishta, Jarrad Koppen, and Marc Agzarian. "Convolutional neural networks for brain tumor segmentation." *Insights into Imaging* 11 (2020): 1-9.
- [54] Sharif, Muhammad Irfan, et al. "Active deep neural network features selection for segmentation and recognition of brain tumors using MRI images." *Pattern Recognition Letters* 129 (2020): 181-189.
- [55] Khan, Muhammad A., et al. "Brain tumor detection and classification: A framework of marker-based watershed algorithm and multilevel priority features selection." *Microscopy research and technique* 82.6 (2019): 909-922.
- [56] <https://www.dicomstandard.org/>
- [57] <https://www.med.upenn.edu/cbica/brats2020/data.html>
- [58] <http://www.miccai.org/>
- [59] <https://brainweb.bic.mni.mcgill.ca/>
- [60] <http://www.med.harvard.edu/aanlib/>
- [61] <https://mail.nmr.mgh.harvard.edu/mailman/listinfo/ibsr>
- [62] <https://neurosynth.org/>
- [63] <http://preprocessed-connectomes-project.org/abide/>
- [64] <https://biosciencedbc.jp/en/>,
- [65] [https://pgimer.edu.in/PGIMER\\_PORTAL/PGIMERPORTAL/GlobalPages/JSP/Page\\_Data.jsp?dep\\_id=64](https://pgimer.edu.in/PGIMER_PORTAL/PGIMERPORTAL/GlobalPages/JSP/Page_Data.jsp?dep_id=64).
- [66] <https://open.fda.gov/data/spl/>
- [67] Jia, Y.; Shelhamer, E.; Donahue, J.; Karayev, S.; Long, J.; Girshick, R.; Guadarrama, S.; Darrell, T. Caffe: Convolutional Architecture for Fast Feature Embedding. *arXiv 2014*, arXiv:1408.5093.]
- [68] Abadi, M.; Agarwal, A.; Barham, P.; Brevdo, E.; Chen, Z.; Citro, C.; Corrado, G.S.; Davis, A.; Dean, J.; Devin, M.; et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. *arXiv 2016*, arXiv:1603.04467.
- [69] Team, T.T.D.; Al-Rfou, R.; Alain, G.; Almahairi, A.; Angermueller, C.; Bahdanau, D.; Bastien, F.; Bayer, J.; Belikov, A.; Belopolsky, A.; et al. Theano: A Python Framework for Fast Computation of Mathematical Expressions. *arXiv 2016*, arXiv:1605.02688
- [70] Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. *arXiv 2019*, arXiv:1912.01703
- [71] Goodfellow, I.J.; Warde-Farley, D.; Lamblin, P.; Dumoulin, V.; Mirza, M.; Pascanu, R.; Bergstra, J.; Bastien, F.; Bengio, Y. Pylearn2: A Machine Learning Research Library. *arXiv 2013*, arXiv:1308.4214.
- [72] Chollet, F. Keras: The Python Deep Learning API. 2020. Available online: <https://keras.io/> (accessed on 1 June 2020).
- [73] <https://github.com/Lasagne/Lasagne>
- [74] Vamsidhar, E., P. Jhansi Rani, and K. Rajesh Babu. "Plant disease identification and classification using image processing." *Int. J. Eng. Adv. Technol* 8.3 (2019): 442-446.
- [75] Sai Deepthi et al. "Comparative Analysis of Brain Tumour Detection Using Deep Learning Methods." *International Journal of Scientific & Technology Research* 8.12 (2019).
- [76] Indira, et al. "An Effective Brain Tumor Detection from T1w MR Images Using Active Contour Segmentation Techniques." *Journal of Physics: Conference Series*. Vol. 1804. No. 1. IOP Publishing, 2021.
- [77] Sairam, et al. "CNN Fusion Based Brain Tumor Detection from MRI images using Active Contour Segmentation Techniques." *Journal of Physics: Conference Series*. Vol. 1804. No. 1. IOP Publishing, 2021.
- [78] P. V. Naganjaneyulu, and K. Satya Prasad. "Comparative Analysis of Active Contour Models for Brain Tumor segmentation from T1w MRI Images." 2021 International Conference on Computer Communication and Informatics (ICCCI). IEEE, 2021.

# Fish Species Classification using Optimized Deep Learning Model

J.M. Jini Mol<sup>1</sup>

Research Scholar, Department of Computer Science,  
Malankara Catholic College Mariagiri  
Affiliated to Manonmaniam Sundaranar University  
Tirunelveli, Tamilnadu, India

Dr. S. Albin Jose<sup>2</sup>

Research Supervisor, Department of Computer  
Science, Malankara Catholic College Mariagiri  
Affiliated to Manonmaniam Sundaranar University  
Tirunelveli, Tamilnadu, India

**Abstract**—Classification of fish species in aquatic pictures is a growing field of research for researchers and image processing experts. Classification of fish species in aquatic images is critical for fish analytical purposes, such as ecological auditing balance, observing fish populations, and saving threatened animals. However, ocean water scattering and absorption of light result in dim and low contrast pictures, making fish classification laborious and challenging. This paper presents an efficient scheme of fish classification, which helps the biologist understand varieties of fish and their surroundings. This proposed system used an improved deep learning-based auto encoder decoder method for fish classification. Optimal feature selection is a major issue with deep learning models generally. To solve this problem efficiently, an enhanced grey wolf optimization technique (EGWO) has been introduced in this study. The accuracy of the classification system for aquatic fish species depends on the essential texture features. Accordingly, in this study, the proposed EGWO has selected the most optimal texture features from the features extracted by the auto encoder. Finally, to prove the efficacy of the proposed method, it is compared to existing deep learning models such as AlexNet, Res Net, VGG Net, and CNN. The proposed method is analysed by varying iterations, batches, and fully connected layers. The analysis of performance criteria such as accuracy, sensitivity, specificity, precision, and F1 score reveals that AED-EGWO gives superior performance.

**Keywords**—Fish species classification; deep learning; GW optimization; auto encoder decoder; feature selection

## I. INTRODUCTION

Object classification is an important area of research for underwater environments. To perform this, a high-resolution camera is used to scatter light and its absorption nature underwater [1]. Numerous researchers are interested in analyzing the health status of aquatic organisms, specifically the population and distribution of fish species [2]. Warming of the oceans will weaken aquatic life by increasing the pressure on fish species [3]. Accordingly, a cost-effective approach must be designed for underwater fish species analysis. In the past, fish species were classified by a laborious process involving the capture of fish or visual surveys conducted by deep-sea divers. Low contrast in the aquatic environment leads in very blurry pictures [4]. Due to the low quality of the images captured underwater, several minute features are lost. This will

certainly impact the performance of the underwater image analysis system.

With the advent of powerful graphical processing units (GPU) and massive amounts of data, deep learning algorithms have become popular in classification and pattern reorganization [5][6]. This study's primary objective is to develop the deep learning model in order to create a completely automated system for classifying fish species. However, the presence of noise in underwater images limits the deep learning models training capacity. Additionally, it makes deep learning models more computationally demanding. This study employs the popular deep learning architecture auto encoder-decoder to obtain texture features from underwater images. Feature selection and hyper parameter (learning rate selection, weight updating process, and others) tuning is the most challenging aspect of building deep learning models.

GWO is an evolutionary optimization technique based on grey wolves' hunting mechanism and leadership hierarchy. Comparing Genetic Algorithms (GA) and particle swarm optimization (PSP), numerous studies have demonstrated that the performance of the GWO optimization method is superior. Unfortunately, traditional GWO requires more iterations to determine the optimal value when the data size and image noise rise. So in this research, the existing GWO algorithm has been enhanced to discover the optimal features using the newly introduced EGWO algorithm with fewer iterations.

Finally, three types of experiments have been conducted to prove the efficiency of the proposed fish species classification system. First, the fish species classification efficiency has been evaluated with the most essential parameters, such as accuracy, recall, specificity, precision, and F1-Score. Secondly, the training and validation efficiency of the proposed AED-based deep learning model has been evaluated. Finally, the computational efficiency of the proposed method has been evaluated. Through these three types of experimental analysis, a comparative study is conducted between the proposed method and existing deep learning algorithms such as AlexNet, ResNet, VGGNet, and CNN. On the basis of these three types of experimental observations, it has been proven that the proposed methodology has excellent training efficiency, high accuracy, and low computing overhead. This study's significant contributions are summarized below.

- 1) First, the R, G, and B channels of the underwater images are normalized to enhance the object visibility.
- 2) Second, the fish morphology localization method has been implemented to eliminate objects that do not have an impact on classification.
- 3) Next, an auto encoder-decoder deep learning model is used to extract underwater images' texture and color features.
- 4) Finally, the EGWO approach is introduced to improve feature selection efficiency.

Highlights of the proposed methodology are as follows: In the second section of this study, the previously developed computer-based classification system for fish species is examined in greater detail. The proposed auto encoder-decoder and enhanced grey wolf optimization are discussed in detail in the third section. In the fourth section, the experimental analysis of the proposed fish species classification system has been comparatively analyzed with existing deep learning algorithms. Finally, the proposed method is concluded.

## II. LITERATURE REVIEW

Classification of underwater images is a challenging task. Manual classification methods demand considerable time and effort. In underwater classification, image size, color, texture, inter-class similarity, and intra-class dissimilarity pose the greatest obstacle. Recently, researchers have developed several machine learning and deep learning methods for classifying underwater fish species [7]. This section reviews the existing fish species classification methods.

The classification is mainly carried out on dead sections depending on shape and texture [8, 9]. The fish's length, width, and thickness are identified using laser light [10]. Classification of fish species is very difficult due to variation in luminosity, background, turbidity of water. Moreover, the similarity of shape and color of various fish species is very difficult to classify.

The fishes are categorized depending on the shape and texture patterns in unconstrained nature [11]. Classification of fish species depends on the biomass content [12]. Classification and identification of fish species depend on morphology, texture and geometry [13]. Fish species identification is done using live fish in the open sea [14].

Recognition of fish species can be done from low resolution images [15]. Combining sparse representation and PCA for classification will provide an accuracy of 82.8%. Gaussian mixture combined with support vector machine will provide a recognition rate of 78%. Few conventional methods are done at a constrained manner. Classification of fish species using shape and color of dead fish sections in organized background.

Grouping of fish species is done based on texture and physical features in unrestrained environments [16] [17]. These techniques will provide good results in fish species with the big difference in texture. The biometric approach [18] takes fish species from various distances and different illuminations. Weber's local descriptor provides classification by the Adaboost classifier.

Classification techniques depending on a mixture of various feature extracting approaches and clustering algorithms with input classifiers are less time-consuming and cheaper [19]. Regions with Convolutional Neural Networks (RCNN) provides good accuracy of 82% in fish species classification [20]. PCA method combined with binary hashing function along with SVM classifier is used for fish species recognition and provides an accuracy of 98% [21].

AlexNet model is pre-trained to perform fish species classification [22]. The CNN-based model and transfer learning use Res Net-152, and SVM classifier provides an accuracy of 95% [23]. Classification of fish species by combination of YOLO deep neural network with Gaussian mixture provides good accuracies in two data sets. VGGNet along with convolutional layers, provides fish species classification [24].

A convolutional neural network is built to rapidly classify the behaviorist of fishes. Few pressure settings are completed in the research center, the fishes' behavior positions are acknowledged, and the model database is recognized [25]. To order the fishes, convolutional neural systems are used. The Faster Regional Convolutional Neural Network strategy is used to eliminate the high spot of images [26]. A Deep CNN Fondest is used for fish recognition, and localization and grouping are done using visual data got from cameras [27].

A Mask R-CNN, together with GOTURN is used for real-time applications of fish recognition and classification [28]. A fish grouping using transfer learning and Matlab is used as the primary step of undertaking the issue. FishNet is an adjustment form of AlexNet to categorize different varieties of fishes [29]. An automatic fish classification based on sonar videos classify fishes based on shape and Movement [30].

The limitations of these surveyed publications are summarized below.

- Due to the poor contrast of underwater photographs, a number of existing methods are semi-automated.
- In many existing systems, the classification algorithm is trained on dead underwater images. Clearly, these techniques cannot be utilized to classify fish species in real-time.
- The accuracy of the fish species classification system depends on the color and texture features of the underwater images. However, no specific optimization approaches have been developed in existing methods for selecting the appropriate color and texture features from underwater images.

## III. PROPOSED METHODOLOGY

This section discusses the process flow and methodologies of the proposed fish species classification system. The entire process flow of the proposed methodology is illustrated in Fig. 1.

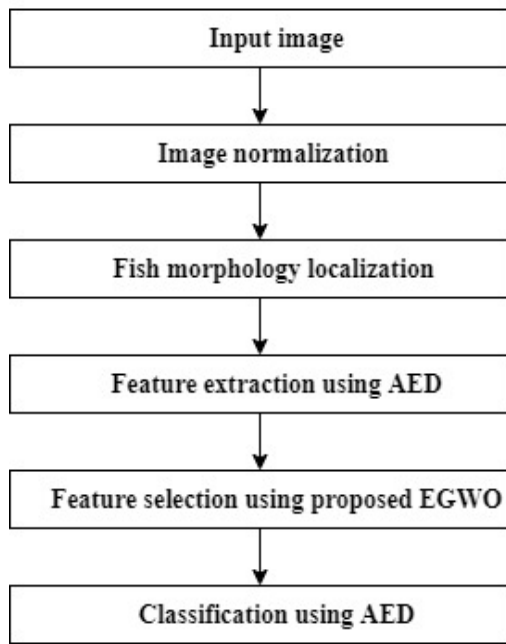


Fig. 1. Overall Process Flow of the Proposed System.

#### A. Dataset

In this study, the dataset is downloaded from GitHub ([https://github.com/primepake/Fishes\\_classification](https://github.com/primepake/Fishes_classification)) and internet sources. There are 1000 images of aquatic fish in this data set. Five type's fish species are included in the data set: aulonocara fire, discus, flame fish, king fish, and molly. Each fish species in the dataset is partitioned into a 4:1 ratio for training and validation purposes. Table I demonstrates the partitioning information for the dataset. Accordingly, 800 images are obtained for the training phase, and 200 are obtained for the testing phase. To boost training efficiency, fish images are artificially augmented using the following image processing techniques: image rotation (90, 180, and 360 degrees), horizontal and vertical flipping, and zooming.

#### B. Image Normalization

Table I demonstrates very clearly that underwater images contain a great deal of noise, particularly undesired dark regions and water backgrounds around the images. These undesirable dark regions impose an additional computational overhead on the classification algorithms. Hence it necessitates the implementation of pre-processing techniques. Generally, all images captured by underwater cameras are RGB images. These images have different color combinations. The majority of images captured from the bottom of the water are quite dark, and the visibility of the objects within them is extremely poor. At the same time, images captured from the surface of the water are extremely bright. When training the model with such images, the accuracy will undoubtedly suffer. Due to the un-normalized nature of underwater images, the red, blue, and green channels should be normalized. Therefore, in this research, the three colour channels of the underwater images are normalized to reduce classification loss. The following equation is used to carry out the underwater image normalization procedure.

$$r(\alpha, \beta) = \frac{R(\alpha, \beta)}{R(\alpha, \beta) + G(\alpha, \beta) + B(\alpha, \beta)} \quad (1)$$

$$g(\alpha, \beta) = \frac{G(\alpha, \beta)}{R(\alpha, \beta) + G(\alpha, \beta) + B(\alpha, \beta)} \quad (2)$$

$$b(\alpha, \beta) = \frac{B(\alpha, \beta)}{R(\alpha, \beta) + G(\alpha, \beta) + B(\alpha, \beta)} \quad (3)$$

Where  $\alpha$  and  $\beta$  indicates the dark and bright regions in the underwater images. ( $r, g$  and  $b$ ) specifies the red, green and blue colour channels of the underwater images. The range  $r(\alpha, \beta), g(\alpha, \beta), b(\alpha, \beta)$  is from 0 to 255, the value of  $r(\alpha, \beta), g(\alpha, \beta), b(\alpha, \beta)$  can vary from 0 to 1. Fig. 12(a) shows the underwater images before the image normalization. Fig. 12(b) shows the underwater images after image normalization.

#### C. Fish Morphology Localization

Fish morphology localization is a crucial element in fish classification. This will eliminate pixels that have no impact on fish classification. This study uses the Simple Linear Iterative Clustering (SLIC) approach to localize fish morphology from aquatic images. SLIC is the most used super pixels segmentation method, and its key benefits include it separates fish regions from aquatic images with little computational cost. SLIC method integrates image plan space and color dimensions to build consistent and realistic super pixels. In order to perform local clustering, the SLIC method performs clustering the pixel dimensions, which is the CIELAB color space ( $l^*a^*b^*$ ). Euclidean distances in  $l^*a^*b^*$  are measured by the following formulas, which calculate the pixel variation in the images.

$$D_{LAB} = \sqrt{((L_i - L_j)^2 + (A_i - A_j)^2 + (B_i - B_j)^2)}, \quad (4)$$

$$D_{xy} = \sqrt{((x_i - x_j)^2 + ((x_i - y_j)^2)}, \quad (5)$$

$$D = D_{LAB} + \frac{m}{s} D_{xy} \quad (6)$$

L=Lightness.

A=Red/Green Values.

B=Blue/Yellow Values.






The SLIC approach begins with cluster sample centers that are uniformly separated. The centers are then moved to initialization places based on the gradient position with the lowest value. Gradient values are calculated by the following formula (7).

$$g(x, y) = ||L(x + 1, y) - L(x - 1, y)||^2 + ||L(x, y + 1) - L(x, y - 1)||^2 \quad (7)$$

$L(x, y)$  is the pixel coordinates. According to the SLIC methodology, pixels in the neighbourhood of a large section will have the same labeling. The method then establishes relationships by relabeling disconnected portions with the labels of the closest neighboring cluster. Fig. 11(c) shows the proposed fish morphology localization procedure.



TABLE I. FISH DATASET DETAILS

Fish image	Total number of images	Training	Testing	Class Label
	200	140	60	King fish
	200	140	60	Discus
	200	140	60	Flame fish
	200	140	60	Molly
	200	140	60	Aulonocara fire

#### D. Auto Encoder Decoder

An auto encoder decoder is a deep learning method that uses unsupervised learning to conduct encoding and decoding [31][32]. Similar to an artificial neural network, it consists of input, hidden, and output layers [33]. Each layer has neurons, with the input and output layers having the same amount, but the hidden layer has fewer neurons than the input layer.

Fig. 2 depicts the architecture of auto encoder decoders. The pre-processed images from the fish dataset are supplied to the auto encoder for feature extraction. There are encoder and decoder sections here. Each encoder comprises a convolution with a filter bank, max pooling, and subsampling to generate the feature map. The encoder is composed of two convolution layers and an intermediate layer. Here, the convolution process of feature maps is not performed. Following batch normalisation, the convergence of local minima is enhanced. Additionally, the decoder comprises two layers of convolution. The convolutional auto encoder consists of output data Y that

generates from input data x that is comparable to the original input data. When the time reaches infinity, the optimal value of the cost function will be reached. F () and F\*(()) are the encode and decode functions. In this study, the nonlinear activation function Rectified Linear Unit (ReLU) is employed, which is represented by the following equation (8). If the function receives negative input, it returns 0, but for positive input, it returns the input value.

$$f(x) = \max(0, x) \quad (8)$$

By optimising the weight and bias term, it is possible to minimise the error. Using the back propagation method, the weight is modified.

$$F(X) = \sigma(wX + B) \quad (9)$$

$$X = \sigma(w(F(X)) + B) \quad (10)$$

The first layer of encoder has an input X. The next layer has input data  $X^L$ , weight  $W^L$  and bias term  $B^L$ . So the above equations are changed as

$$F(X^{L+1}) = \sigma(W^L X^L + B^L) \quad (11)$$

$$X^{L+n+1} = \sigma(W^{n-L}(F(X^{n+L})) + B^{n-L}) \quad (12)$$

This auto encoder is trained for 200 epochs. The weights are optimized in order to find the local minima. Hidden layer takes output of the preceding layer. Throughout the training phase  $\sigma$  value was 4. The mean squared error decreased till 14 for 100 epochs and the performance during the training phase is 0.00058. The error rate is reduced smoothly without over fitting. The trained parameters will be provided to the next layer. Each and every auto encoder is subjected to max polling layer and the result is subsampled by a factor of 2.

$$y_{i,j} = \text{maximum}(x_{i,j}) \quad (13)$$

Where  $x_{i,j}$  and  $y_{i,j}$  denotes the  $i$  region of  $j$  feature-map which is the feature map of input and  $i$  neuron of  $j$  feature-map which is the feature map of output respectively. The number feature-map remain the same both at input and output. In each and every decoder up sampling is done for its feature-map by applying the indices of maximum pooling from the corresponding feature-map of encoder. As a result sparse feature-map is produced. Convolution of decoder filter bank and feature map is done to regenerate the input.

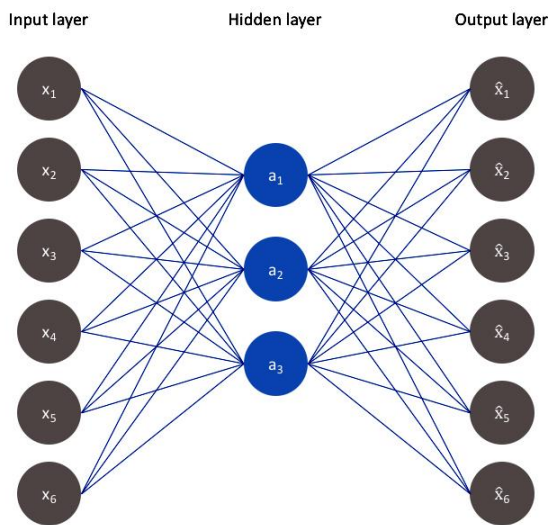


Fig. 2. Auto Encoder Decoder.

### E. Traditional GWO Method

This algorithm is based on the searching and hunting behaviour of grey wolves. According to GWO, the fitness value has three parameters one is alpha ( $\alpha$ ), the other is beta ( $\beta$ ), the third is the delta ( $\delta$ ). The hierarchical structure of grey wolves is illustrated in Fig. 3. The remaining solutions are called as omegas ( $\omega$ ). Three grey wolves will guide omegas in searching step. At time  $t=1$ , first iteration begins, at the time when a prey is found out. The omegas will encircle the prey with the help of alpha, beta and delta wolves. Three coefficients ( $a^{\rightarrow}, c^{\rightarrow}, d^{\rightarrow}$ ) will help in encircling process. They are.

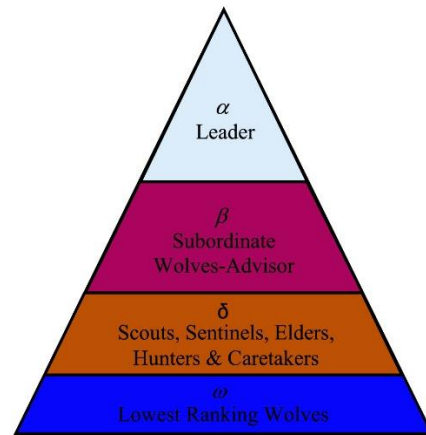


Fig. 3. Hierarchy of Grey Wolves.

$$d_{\alpha}^{\rightarrow} = |c_1^{\rightarrow} \cdot x_{\alpha}^{\rightarrow} - x_{(t)}^{\rightarrow}| \quad (14)$$

$$d_{\beta}^{\rightarrow} = |c_2^{\rightarrow} \cdot x_{\beta}^{\rightarrow} - x_{(t)}^{\rightarrow}| \quad (15)$$

$$d_{\delta}^{\rightarrow} = |c_3^{\rightarrow} \cdot x_{\delta}^{\rightarrow} - x_{(t)}^{\rightarrow}| \quad (16)$$

Where  $t$  is the current iteration.  $x_1^{\rightarrow}, x_2^{\rightarrow}, x_3^{\rightarrow}$  denote the position vector of alpha, beta and delta respectively.

$$x_1^{\rightarrow} = x_{\alpha}^{\rightarrow} - a_1^{\rightarrow} \cdot d_{\alpha}^{\rightarrow} \quad (17)$$

$$x_2^{\rightarrow} = x_{\beta}^{\rightarrow} - a_2^{\rightarrow} \cdot d_{\beta}^{\rightarrow} \quad (18)$$

$$x_3^{\rightarrow} = x_{\delta}^{\rightarrow} - a_3^{\rightarrow} \cdot d_{\delta}^{\rightarrow} \quad (19)$$

$$x_{(t)}^{\rightarrow} = \frac{x_1^{\rightarrow} + x_2^{\rightarrow} + x_3^{\rightarrow}}{3} \quad (20)$$

The parameters  $a^{\rightarrow}, c^{\rightarrow}$  are given by

$$a^{\rightarrow} = 2\alpha r_1^{\rightarrow} - \alpha \quad (21)$$

$$c^{\rightarrow} = 2r_2^{\rightarrow} \quad (22)$$

Where  $r_1^{\rightarrow}$  and  $r_2^{\rightarrow}$  are random numbers. The controlling parameter is  $\alpha$  which alters the value of  $a^{\rightarrow}$ . If the value of  $a^{\rightarrow}$  is greater than 1 the grey wolves will move farther away from the prey, which denotes that the omega will run away, representing global optimization. If the value of  $a^{\rightarrow}$  is lesser than 1, the omega will move towards the prey showing local optimization. The controlling parameter will decline from 2 to zero in a linear manner. This is given by

$$\alpha = 2\left(1 - \frac{it}{n}\right) \quad (23)$$

Where  $n$  is the maximum value of iteration number which is a cumulative iteration number.

Every grey wolf runs away or moves towards the prey with a suitable mean weight for an alpha, beta, and delta. On the start of the searching method, the weight of the alpha must be larger than the others i.e. beta and delta. The hierarchy in weight must be such that the weight of alpha is higher than that of beta and delta. Similarly, the weight of beta is greater than that of delta. The alpha wolf is given greater importance than others. The alpha wolf is considered to be nearer to the prey. The alpha wolf governs the searching. Beta and delta have a less important role. If beta or delta finds the best position, it is

transferred to the alpha wolf. In the searching process, the hypothesized prey is encircled, but in hunting, the real prey is encircled. The alpha is the nearest one to the prey than the beta. But the delta is farther away to beta. The omega wolf will alter and give their best positions to these dominants. Initially, the value of alpha is 1 and the value of beta and delta is zero. Finally, the dominants will encircle the prey, since they have the same weight. The weight of alpha has to be reduced, and the weight of beta and delta has to arise due to the cumulative iteration number. The position is updated as

$$x_{(t+1)}^{\rightarrow} = w_1 \bar{x}_1 + w_2 \bar{x}_2 + w_3 \bar{x}_3^{\rightarrow}$$

Where  $w_1 + w_2 + w_3 = 1$  such that  $w_1 \geq w_2 \geq w_3$  (24)

Algorithm 1 explain the process flow of traditional GWO. In the GWO algorithm, the initial population is generated by randomly. In this study, the population refers to the number of texture features. When initialising at random, there is a significant likelihood of irrelevant or redundant features. This raises the number of iterations of the GWO method, which increases the algorithm's running time and computational overhead. Therefore, the traditional GWO method must be enhanced in order to choose the desirable features from the fish data set with fewer iterations and in less time.

---

#### Algorithm 1

**Step1:** Random initialization of the population of grey wolves Xi

**Step2:** Initialise the values of  $\alpha, \beta, \delta$ .

**Step3:** Fitness of search agent is calculated

$x_a^{\rightarrow}$  = search agent in first position

$x_b^{\rightarrow}$  = search agent in second position

$x_g^{\rightarrow}$  = search agent in third position

**Step4:** when ( $t <$  the number of iterations)

**For** each search agent

update the position of current agent using equation (24)

**end for**

**Step5:** Update the values of  $\alpha, \beta, \delta$ .

**Step6:** The fitness of all agents are computed.

**Step7:** Update  $x_a^{\rightarrow}, x_b^{\rightarrow}, x_g^{\rightarrow}$

**Step8:** Increment the value of t

**End while**

**Step9:** Return the value of  $x_a^{\rightarrow}$

---

#### F. Feature Selection using Enhanced GWO Method

Feature selection is a very important part of artificial intelligence based prediction models [33][34]. Fig. 4 shows the

overall process flow of the proposed AED-EGWO. The proposed research AED extracts different types of colour and texture features from underwater images. When extracting features from underwater images using AED, redundant and irrelevant features are likely to be extracted. The formulas 25-33 determine the important texture features of this proposed method. The accuracy of classification algorithms is relies on effective feature selection techniques. The precision of deep learning models can be significantly enhanced by picking the most beneficial feature.

In addition, the computational burden of deep learning models is drastically lowered. The EGWO approach is used to eliminate irrelevant and redundant texture features from this classification system for fish species.

$$\text{Autocorrelation} = \sum_i \sum_j p(i, j) p(i, j) \quad (25)$$

$$\text{Contrast} = \sum_{n=0}^{N_e-1} n^2 \sum_{n=i}^{N_e} \sum_{n=j}^{N_e} \{p(i, j)\} \quad (26)$$

$$\text{Correlation} = \frac{\sum_i \sum_j p(i, j) p(i, j) - \mu_x \mu_y}{\sigma_x \sigma_y} \quad (27)$$

$$\text{Energy} = \sum_i \sum_j p(i, j) \quad (28)$$

$$\text{Dissimilarity} = \sum_i \sum_j |i - j| * p(i, j) \quad (29)$$

$$\text{Entropy} = \sum_i \sum_j p(i, j) \log(p(i, j)) \quad (30)$$

$$\text{Homogeneity} = \sum_i \sum_j \frac{1}{1+(i-j)^2} p(i, j) \quad (31)$$

$$\text{Variance} = \sum_i \sum_j (i - \mu)^2 p(i, j) \quad (32)$$

$$\text{Cluster shade} = \sum_i \sum_j (i + j - \mu_x - \mu_y)^2 p(i, j) \quad (33)$$

Here, an improved version of the GWO algorithm is proposed. Particularly in the population initialization phase, provide an intelligent initialization approach to achieve the optimal solution in early iterations. This intelligent initialization strategy accelerates the convergence of the algorithm. The key difference is that the population is now initialized using a correlation-based technique rather than at random method. The initial population is formed based on the correlation value, which decides whether a feature value is selected or not. The following equation represents the computation of the correlation for a feature f:

$$\text{Cor}_F = \frac{\sum(F_i - \bar{F})(C_i - \bar{C})}{\sqrt{\sum(F_i - \bar{F})^2 \sum(C_i - \bar{C})^2}} \quad (34)$$

Above equation, F indicates the texture features of underwater images. C represents the class values.  $\bar{F}$  and  $\bar{C}$  represents the mean values of features and classes, respectively.

EGWO Methodology for feature selection

AED Features extraction

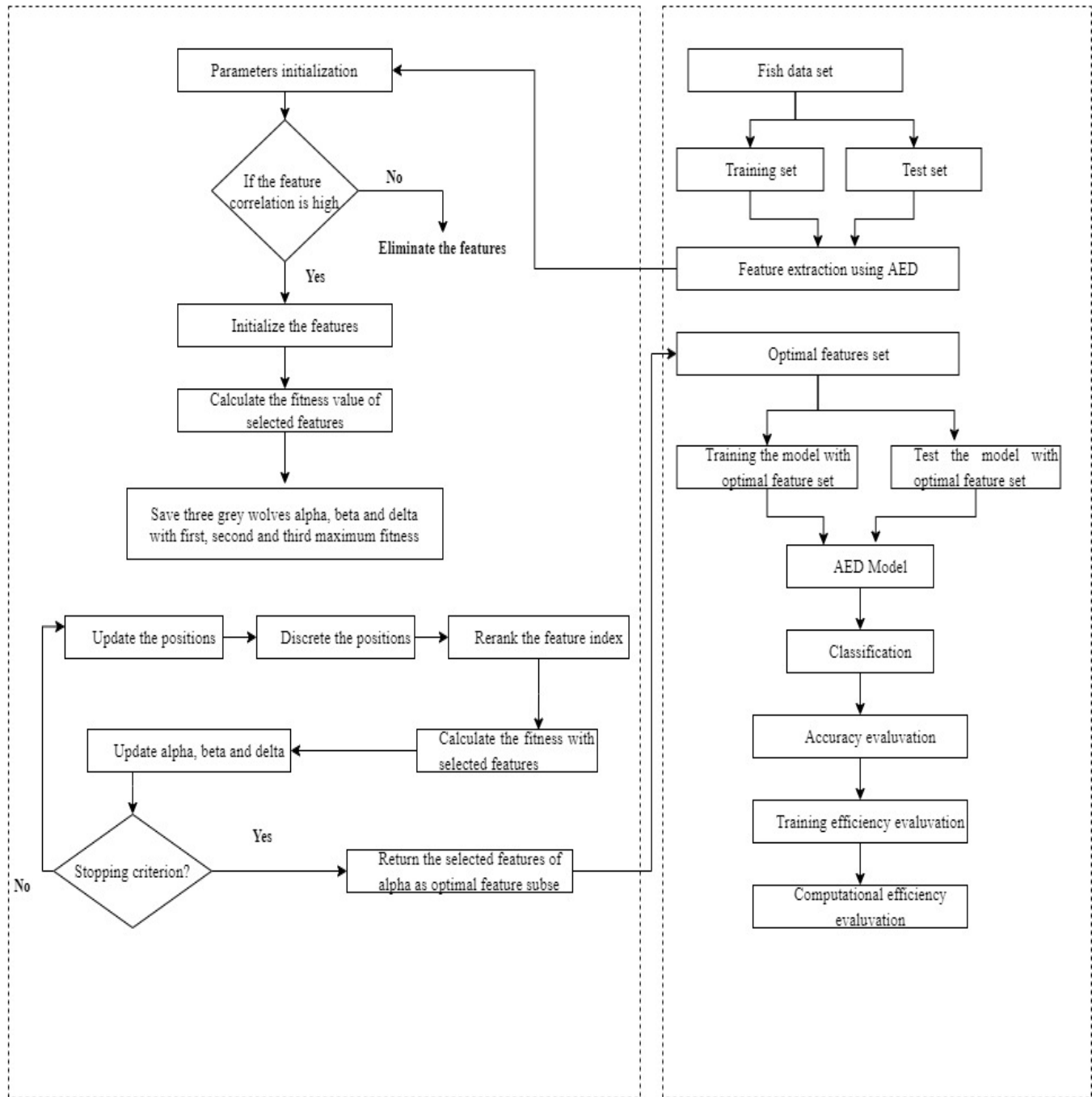


Fig. 4. Overall Process Flow of the Proposed AED-EGWO.

A feature with higher correlation values is crucial for the classification. Therefore, employing the following formula, the suggested method assures that features with high correlation values are included in the initial population. Then, based on the correlation values, the new population is initialized as follows.

$$P(i) = \begin{cases} 1, & \text{if } rand < IG(i) \\ 0, & \text{if } rand \geq IG(i) \end{cases} \quad (35)$$

*rand* is a random number between 0 and 1, and  $P(i)$  is the binary representation of the *i*th feature in the initial population. According to formula (35), the features values with high correlation values are initialized for the initial population. This allows the traditional GWO method to get optimal features in a minimal number of iterations. Algorithm 2 explain the proposed EGWO based texture feature selection.

---

**Algorithm 2 proposed feature selection**

---

Input: AED extracts texture features from fish training data set

$(T_1, T_2, T_3 \dots T_n, T_c)$   $T_n$  is the number of texture features and  $T_c$  is the target class.

Output:  $O_{tf}$  (Optimal texture features)

Step 1: **Begin**

Step 2 : **for**  $i=1$  to  $n$  **do**

C=calculate the coefficient( $T_n, T_c$ ) using formula

**End**

Set the threshold level  $\tau=0.6$

Step 3: **if**  $\tau < 0.6$  // means there is no significant correlation between  $T_n$  and  $T_c$

Step 4: **for**  $I=1$  to  $m$  **do**

r=Calculate the significant between (C, $\tau$ )

**if** significant is high

**Then**

Add the features to  $\Rightarrow O_{tf}$

**End**

**End**

**Return**  $O_{tf}$

**End**

---

#### IV. RESULTS AND DISCUSSION

##### A. System setup and Configuration

The software tools used to develop this fish species categorization system are Matlab 2018 and deep learning libraries. Additionally, Windows 10 is used as an operating system. To run deep learning libraries and design the implementation model for this fish species classification system, a graphics processing unit with 4GB NVIDIA 1650, an Intel 10th Gen Core i5 processor, 256GB SSD, and 16GB RAM is utilized.

##### B. Classification Model Performances Evaluation

Existing deep learning approaches such as AlexNet, ResNet, VGGNet, and CNN are compared to the proposed methodology. The proposed method is evaluated with different configurations of the deep learning model, including fully connected layers, iterations, and batches, with and without optimization. The test image and the ground truth image must be compared. Five common performance measures that are typically used for classification are used to the analysis the performance of the proposed classification system: accuracy, sensitivity, specificity, precision, and F1 score. The mentioned accuracy measures are depending on the following variables.

**True positive fish species classification:** If the proposed method correctly recognises and classifies fish species from underwater images, this classification is known as a true positive fish species classification. The variable TP specifies the classification of true positive fish species.

**True negative fish species classification:** If the proposed method correctly identifies and classifies non-fish species from underwater images, this classification is known as true negative

fish species classification. The variable TN specifies classification of true negative fish species.

**False positive fish species classification:** If the proposed method mistakenly identifies and classifies non-fish species as fish species from underwater images, this is referred to as a false positive fish species classification. The variable FP specifies the classification of false fish species.

**False negative fish species classification:** False negative fish species classification happens when the proposed approach fails to recognise and classify non-fish species from underwater images. The variable FN specifies classification of false negative fish species.

The confusion matrices of the proposed and existing deep learning models are depicted in Fig. 5. According to Table I, 300 images are utilized to test the AED-EGWO approach. Accuracy, recall, specificity, precision, and F1-Score values are calculated based on the Confusion matrix, which are shown in Fig. 6 to 10.

In this study, the definition of accuracy is the correct classification of fish species from underwater images. To ensure the reliability of proposed model, it is essential to determine the proportion of true positives and true negatives among all of the instances that have been analyzed. Mathematically, accuracy is expressed as:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (36)$$

According to the classification of fish species, sensitivity involves accurately identifying fish species. Sensitivity can be determined by examining the proportion of true positives. Sensitivity has the following mathematical expression:

$$Sensitivity (Recall) = \frac{TP}{TP+FN} \quad (37)$$

According to the classification of fish species, specificity determines the reliability of non-fish classification results. To measure it, calculate the proportion of genuine negatives. Specificity has the following mathematical expression:

$$Specificity = \frac{TN}{TN+FP} \quad (38)$$

Precision is the ratio of the number of accurate fish classifications to the total number of positive fish predictions. Precision is calculated by the following formula.

$$Precision = \frac{TP}{TP+FP} \quad (39)$$

In an AI-based classification system, the F1-score is calculated using the overall precision and recall values. Formula for calculating the F1-score is as follows.

$$F1 - score = \frac{2(Recall \times Precision)}{Recall + Precision} \quad (40)$$

The confusion matrices of the proposed and existing deep learning models are depicted in Fig. 5. Accuracy, recall, specificity, precision, and F1-Score values are calculated based on the Confusion matrix, which are shown in Fig. 6 to 10.



Fig. 5. TP, TN, FP and FN Ratios of Proposed and Existing Methods.



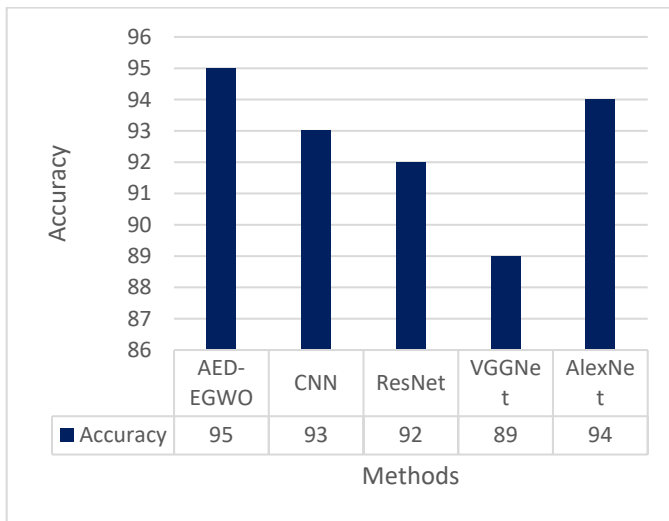


Fig. 6. Accuracy Comparison Results.

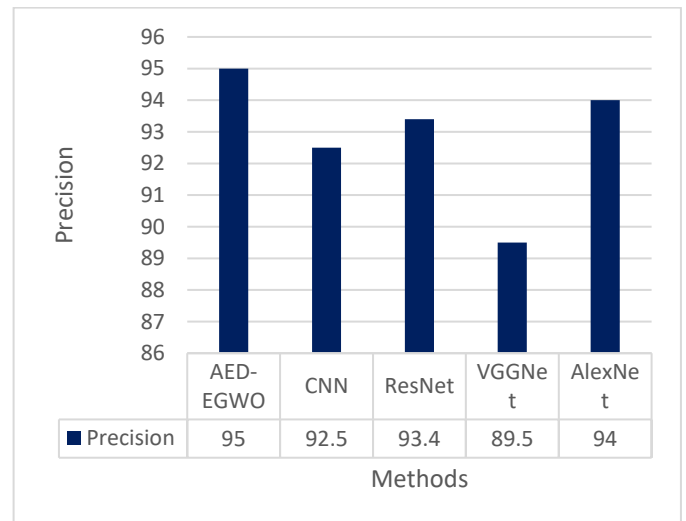


Fig. 9. Precision Comparison Results.

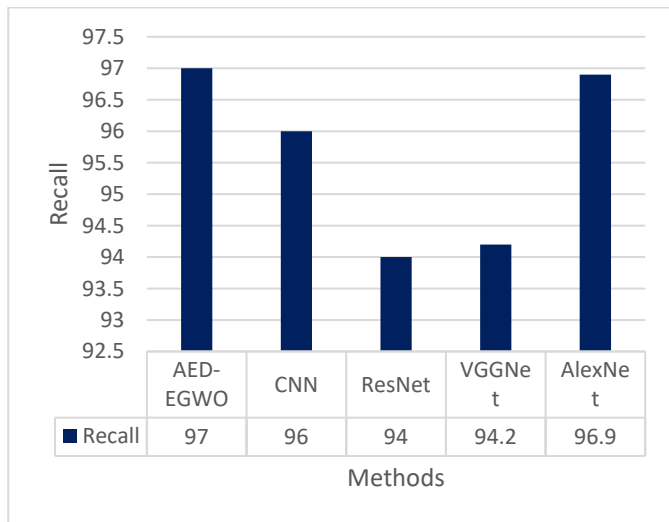


Fig. 7. Recall Comparison Results.

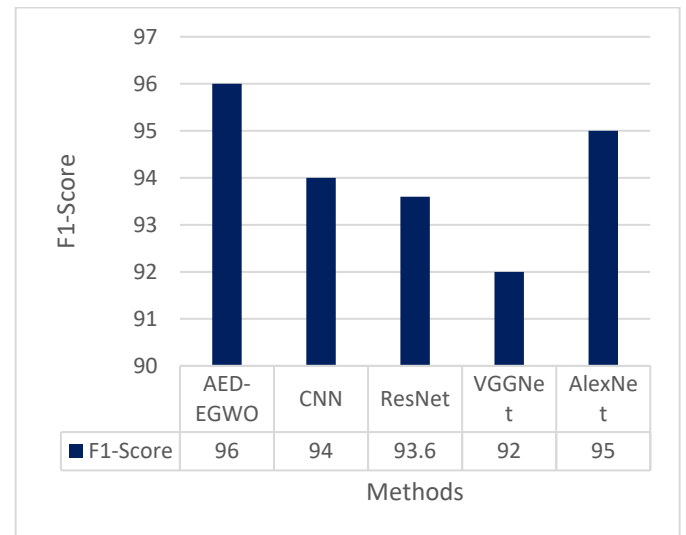


Fig. 10. F1-score Comparison Results.

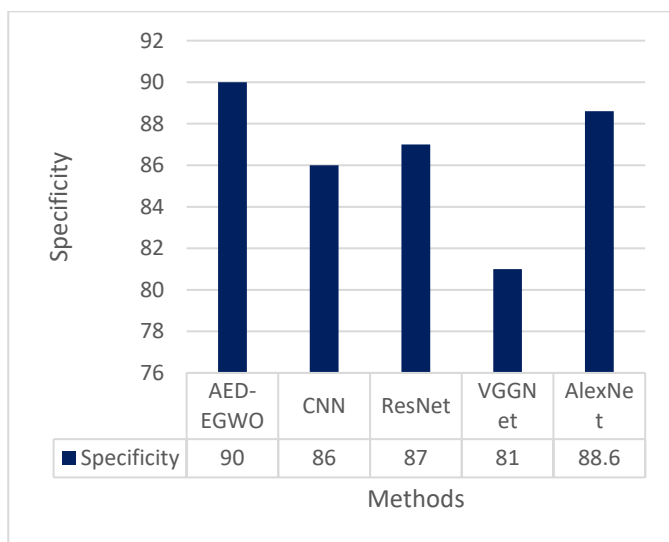


Fig. 8. Specificity Comparison Results.

The proposed methodology's performance is compared to traditional methods AlexNet, ResNet, VGGNet, and CNN. For a better analysis, the methods are experimentally implemented and their performance metrics are compared with the existing methods. The experimental results (accuracy, recall, specificity, precision and F1-Score) are presented in the Fig. 6 to 10. According to the experimental results, it is discovered that the proposed method is more accurate than the other state-of-the-art methods. Proposed fish classification system will have a greater number of true-positive fish pixel classifications; this will obviously improve the accuracy metrics. Following the proposed system, CNN and AlexNet have the highest accuracy (accuracy, recall, specificity, precision, and F1-Score). The ROC curve depicted in Fig. 10 has an AUC between 0.5 and 1.0, which indicates that it ranks a random positive sample higher than a random negative sample more than fifty percent of the time. Based on the results, it can be seen that the proposed methodology has fewer false positive fish pixels than other techniques. In addition, it is recognized that the number of true negative fish pixels is greater for the

proposed method than for the traditional deep learning techniques. In addition, the experimental results proves suggested technique reduces the number of false negative fish pixels.

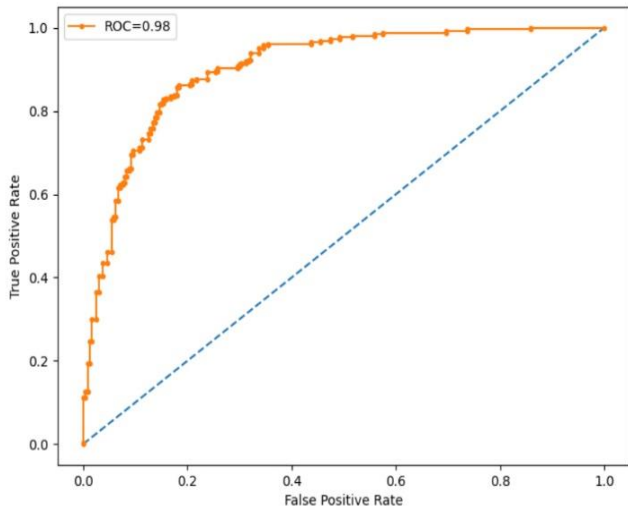


Fig. 11. ROC Curve.

### C. Training efficiency analysis

This section examines the training efficiency of the proposed and existing approaches. For that, mean absolute error and mean squared error are two of the most important training efficiency evaluation metrics applied in this research. If the error values are minimal, therefore the training loss of the suggested technique is also limited.

1) *Mean absolute error*: Mean absolute error (MAE) is the average of the difference between actual and predicted values. It describes the variation between the predicted and the actual values. The following formula calculates MAE.

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (41)$$

$y$  represents the actual class value, while  $\hat{y}$  represents the outcome predicted by the proposed model.

2) *Mean squared error*: Mean squared error is the average of the squared differences between the actual and predicted values. The equation below is used to compute the MSE.

$$MSE = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}_i|^2 \quad (42)$$

MAE and MSE provide positive integer values during training. If the value is near to zero, the deep learning model's training loss is very low; otherwise, the training loss is high. The MAE and MSE comparisons are summarized in Table II, it is observed that the proposed system has lower MAE and MSE. The proposed system is to less error-prone than existing traditional methods. As the error value is lower for the proposed approach, it is expected that the proposed system will be more effective than other conventional deep learning methods.

### D. Computational Efficiency Analysis

Table III summarized the computational efficiency of the AED-EGWO and existing methods. According to the experimental data, AED-EGWO takes 16 minutes to train the model, which is the shortest training time in the table, while RsetNet takes the longest of 23 minutes.

### E. Discussion

Classification of fish species is an essential component of marine research and oceanography. It contributes significantly to the migration, breeding, and monitoring of endangered fish species. In the meantime, the manual classification of fish species is a labour-intensive and time-consuming process. To automate the classification procedure, numerous computer-aided fish species classification systems have been developed. In existing methods, the deep learning or machine learning models are trained using images of dead fish taken out from under water. When training deep learning models using these photos instead of real-time underwater photographs, the FN and FP rates are increased. High FN and FP rates have a noticeable impact on the classification model's precision. To prevent this, the AED model in this study was trained using underwater photos of live fish. However, developing fish classification system using deep learning models with photos of underwater fish presents numerous challenges. Especially underwater, light penetration is very low, therefore pictures captured from this environment are extremely dim. Accordingly, the visibility of objects in underwater photographs are very poor. Therefore, underwater images have been normalized in this research to correct the problem. The normalized images through this research are shown in Fig. 12(b). Also, in the images taken from underwater, the color of water and other objects besides fish are occupied excessively, which increases the computational burden of the deep learning model. To correct it, in this research, the morphology of fish has been localized using the Simple Linear Iterative Clustering (SLIC) method. The results of fish morphology localization are shown in Fig. 12(b). Further, this study proposes the AED-EGWO framework for the classification of aquatic fish species. EGWO is being proposed for two significant responsibilities here.

TABLE II. TRAINING EFFICIENCY COMPARISON RESULTS

Methods	MAE	MSE
AED-EGWO	0.23	0.0529
CNN	0.42	0.1764
ResNet	0.77	0.5929
VGGNet	0.47	0.2209
AlexNet	0.31	0.0961

TABLE III. COMPUTATIONAL EFFICIENCY COMPARISON RESULTS

Methods	Training time	Classification time
AED-EGWO	16 minutes	1.37 Seconds
CNN	21 minutes	2.18 Seconds
ResNet	23 minutes	2.71 Seconds
VGGNet	19 minutes	1.92 Seconds
AlexNet	17 minutes	1.68 Seconds

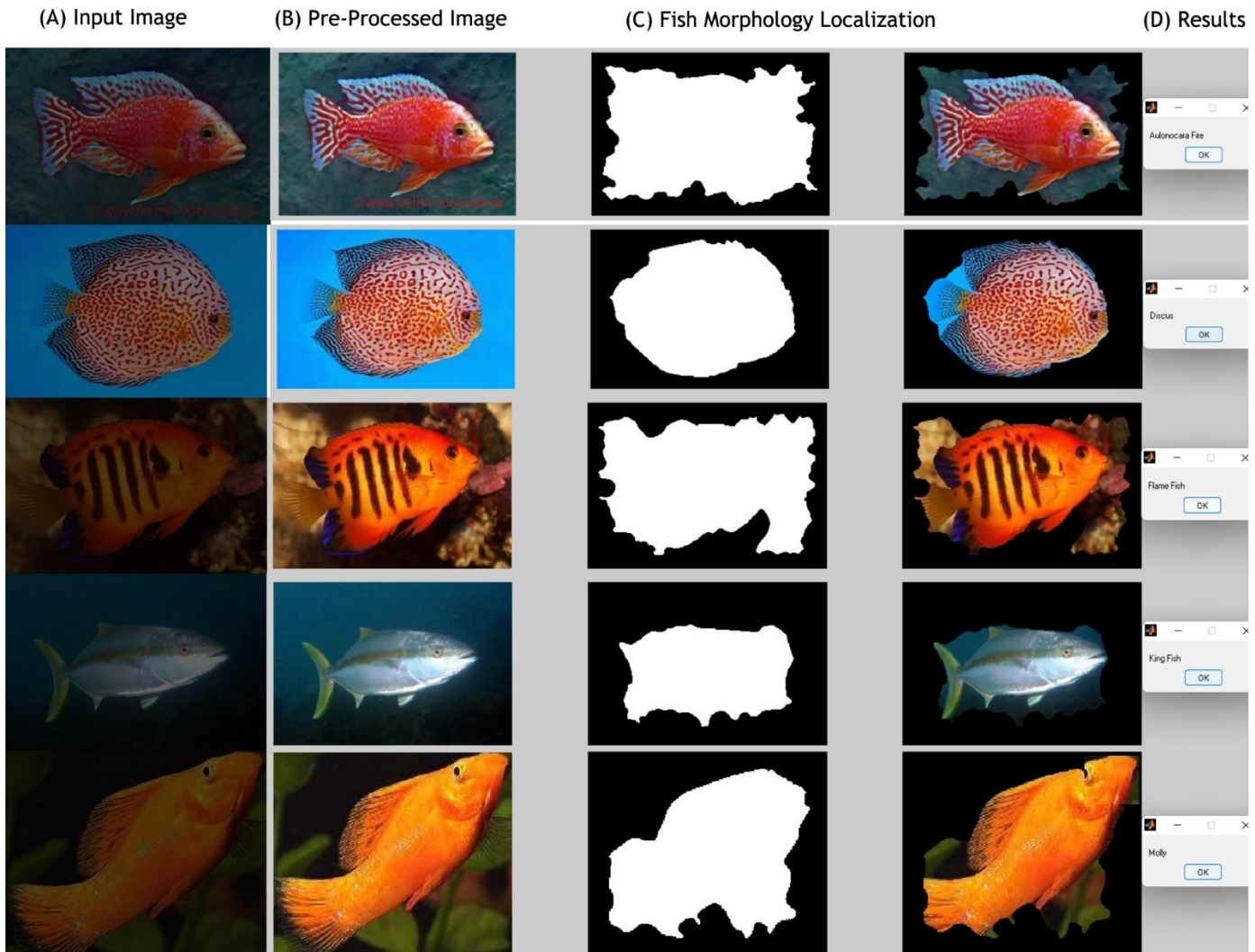


Fig. 12. Experimental Results.

Initially, EGWO eliminates irrelevant and redundant texture information from the data set during AED model training. Next, EGWO method select the optimal texture features from the data set with the lowest possible iterations.

According to the proposed EGWO, feature initialization is based on correlation rather than randomness. Therefore, it only initializes features with a strong relationship to classification results. Thus, the optimization method can find optimal features with less iterations. This eliminates the unnecessary processing resources required for AED training. Further, it enhances the fish classification accuracy to some extent.

## V. CONCLUSION

In this study, the AED-EGWO methodology is developed for classifying fish species. This recommended classification method has two major components: optimal feature selection using EGWO and fish classification using AED. First, an improved grey wolf optimization approach is designed to identify the most important texture feature in the fish data set with the fewest possible iterations. Second, an auto encoder decoder network is developed to classify fish species based on

the identified features. Finally, experimental study has been conducted to evaluate the proposed method's reliability and classification effectiveness. The experimental results of the proposed AED-EGWO approach were compared to existing deep learning models. The comparison results demonstrate that the proposed strategy has greater classification precision and reduced training loss.

This fish classification system was evaluated using images of clear sea water. In the future, this research can be expanded to classify fish species in blurry fresh water images.

## REFERENCES

- [1] D. Rathi, S. Jain and S. Indu, "Underwater Fish Species Classification using Convolutional Neural Network and Deep Learning," 2017 Ninth International Conference on Advances in Pattern Recognition (ICAPR), 2017, pp. 1-6, doi: 10.1109/ICAPR.2017.8593044.
- [2] F. J. P. Montalbo and A. A. Hernandez, "Classification of Fish Species with Augmented Data using Deep Convolutional Neural Network," 2019 IEEE 9th International Conference on System Engineering and Technology (ICSET), 2019, pp. 396-401, doi: 10.1109/ICSEngT.2019.8906433.
- [3] S. Hasija, M. J. Buragohain and S. Indu, "Fish Species Classification Using Graph Embedding Discriminant Analysis," 2017 International

- Conference on Machine Vision and Information Technology (CMVIT), 2017, pp. 81–86, doi: 10.1109/CMVIT.2017.23.
- [4] M. T. A. Rodrigues, F. L. C. Pádua, R. M. Gomes and G. E. Soares, "Automatic fish species classification based on robust feature extraction techniques and artificial immune systems," 2010 IEEE Fifth International Conference on Bio-Inspired Computing: Theories and Applications (BIC-TA), 2010, pp. 1518–1525, doi: 10.1109/BICTA.2010.5645273.
- [5] Sarker IH. Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions. SN Comput Sci. 2021;2(6):420. doi: 10.1007/s42979-021-00815-1. Epub 2021 Aug 18. PMID: 34426802; PMCID: PMC8372231. FLEX Chip Signal Processor(MC68175/D), Motorola, 15(3) (1996) 250–275.
- [6] Alzubaidi, L., Zhang, J., Humaidi, A.J. et al. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. J Big Data 8, 53 (2021). <https://doi.org/10.1186/s40537-021-00444-8>.
- [7] Fouad, M. M., Zawbaa, H. M., Gaber, T., Snasel, V., & Hassanien, A. E. (2016). A fish detection approach based on BAT algorithm. In The 1st international conference on advanced intelligent system and informatics, AISI 2015 (pp. 273–283). Cham: Springer.
- [8] Spampinato, C., Giordano, D., Di Salvo, R., Chen-Burger, Y. H. J., Fisher, R. B., & Nadarajan, G., (2010). Automatic fish classification for underwater species behavior understanding. In Proceedings of the first ACM international workshop on analysis and retrieval of tracked events and motion in imagery streams (pp. 45–50). ACM.
- [9] Nagashima, Y., & Ishimatsu, T. (1998). A morphological approach to fish discrimination. In IAPR workshop on machine vision applications, Nov. 17–19 (pp. 306–309).
- [10] Storbeck, F., & Daan, B. (2001). Fish species recognition using computer vision and a neural network. Fisheries Research, 51(1), 11–15.
- [11] Rova, A., Mori, G., & Dill, L. M. (2007). One fish, two fish, butterfly, trumpeter: Recognizing fish in underwater video. In IAPR conference on machine vision applications, Tokyo, Japan (pp. 404–407).
- [12] Shafait, F., Mian, A., Shortis, M., Ghanem, B., Culverhouse, P. F., Edgington, D., et al. (2016). Fish identification from videos captured in uncontrolled underwater environments. ICES Journal of Marine Science, 73(10), 2737–2746.
- [13] Hernández-Serna, A., & Jiménez-Segura, L. F. (2014). Automatic identification of species with neural networks. PeerJ, 2, e563.
- [14] Huang, P. X., Boom, B. J., & Fisher, R. B. (2012). Hierarchical classification for live fish recognition. In BMVC student workshop paper.
- [15] Sun, X., Shi, J., Dong, J., & Wang, X. (2016). Fish recognition from low-resolution underwater images. In 9th International congress on image and signal processing, biomedical engineering and informatics (CISP-BMEI), IEEE (pp. 471–476).
- [16] Ludwig Bothmann, Michael Windmann, Goeran Kauermann, "Realtime classification of fish in underwater sonar videos" 2016.
- [17] [Spampinato C, Giordano D, Salvo RD, Chen-Burger Y-HJ, Fisher RB, Nadarajan G (2010) Automatic fish classification for underwater species behavior understanding. In: Proceedings of the first ACM international workshop on analysis and retrieval of tracked events and motion in imagery streams, pp 45–50.
- [18] Zhao J, Cao Y, Fan D, Cheng M, Li X, Zhang L (2019) Contrast prior and fluid pyramid integration for rgbd salient object detection. In: 2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 3922–3931.
- [19] Rauf HT, Ikram Ullah Lali M, Zahoor S, Shah SZH, Rehman AU, Bukhari SAC (2019) Visual features based automated identification of fish species using deep convolutional neural networks. Comput Electron Agric 167:105075.
- [20] Tharwat A, Hemedan AA, Hassanien AE, Gabel T (2018) A biometric-based model for fish species classification. Fish Res 204:324–336.
- [21] Jin L, Liang H (2017) Deep learning for underwater image recognition in small sample size situations. In: OCEANS 2017 - Aberdeen, pp 1–4.
- [22] Sun, X., Shi, J., Dong, J., & Wang, X. (2016). Fish recognition from low-resolution underwater images. In 9th International congress on image and signal processing, biomedical engineering and informatics (CISP-BMEI), IEEE (pp. 471–476).
- [23] Qin JLH, Li X, Zhang C (2016) Deepfish: accurate underwater live fish recognition with a deep architecture. Neurocomputing 187:49–58 101088:57
- [24] Rova A, Mori G, Dill LM, "One fish, two fish, butterfly, trumpeter: recognizing fish in underwater video", In: IAPR conference on machine vision applications (2007).
- [25] Fangfang Han, Junchaozhu, Bin Liu, "Fish Shoals Behavior Detection based on Convolutional Neural N/Wand Spatiotemporal Information" 2020.
- [26] Aditya Agarwal , Tushar Malani , Gaurav Rawal , Navjeet Anand, Manonmani S, 2020, Underwater Fish Detection, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) Volume 09, Issue 04 (April 2020),
- [27] Christensen, J. H., Mogensen, L. V., Galeazzi, R., & Andersen, J. C. (2019). Detection, Localization and Classification of Fish and Fish Species in Poor Conditions using Convolutional Neural Networks. In Proceedings of 2018 IEEE/OES Autonomous Underwater Vehicle Workshop (AUV) (pp. 1-6).
- [28] J. H. Christensen, L. V. Mogensen, R. Galeazzi and J. C. Andersen, "Detection, Localization and Classification of Fish and Fish Species in Poor Conditions using Convolutional Neural Networks," 2018 IEEE/OES Autonomous Underwater Vehicle Workshop (AUV), 2018, pp. 1-6, doi: 10.1109/AUV.2018.8729798.
- [29] C.S.Arvind, R.Prajwal,Prithvi Narayana Bhat,A.Sreedevi, K.N.Prabhudeva ” , Fish Detection and Tracking in Pisciculture Environment using Deep Instance Segmentation, 2019.
- [30] Suryadiputra Liawatimena, Yaya Heryadi, Lukas, "A Fish Classification on Images using Transfer Learning and Matlab", 2018.
- [31] Boulard, H., Kabil, S.H. Autoencoders reloaded. Biol Cybern 116, 389–406 (2022). <https://doi.org/10.1007/s00422-022-00937-6>.
- [32] Li, Y., Wang, Z., Yang, X. et al. Efficient convolutional hierarchical autoencoder for human motion prediction. Vis Comput 35, 1143–1156 (2019). <https://doi.org/10.1007/s00371-019-01692-9>.
- [33] Qiang Li, Huiling Chen, Hui Huang, Xuehua Zhao, Zhen Nao Cai, Changfei Tong, Wenbin Liu, Xin Tian, "An Enhanced Grey Wolf Optimization Based Feature Selection Wrapped Kernel Extreme Learning Machine for Medical Diagnosis", Computational and Mathematical Methods in Medicine, vol. 2017, Article ID 9512741, 15 pages, 2017. <https://doi.org/10.1155/2017/9512741>.
- [34] Muni, MK, Parhi, DR, Kumar, PB. Implementation of grey wolf optimization controller for multiple humanoid navigation. Comput Anim Virtual Worlds. 2020; 31:e1919. <https://doi.org/10.1002/cav.1919>.

# Environmental Noise Pollution Forecasting using Fuzzy-autoregressive Integrated Moving Average Modelling

Muhammad Shukri Che Lah<sup>1</sup>, Nureize Arbai<sup>2</sup>, Syahir Ajwad Sapuan<sup>3</sup>, Pei-Chun Lin<sup>4</sup>

Faculty of Computer Science and Information Technology

Universiti Tun Hussein Onn Malaysia, 86400 Batu Pahat, Malaysia<sup>1,2,3</sup>

Dept. of Information Engineering and Computer Science, Feng Chia University, No. 100, Wenhwa Rd, Taichung, Taiwan<sup>4</sup>

**Abstract**—Predicting noise pollution from building sites is important to take precautions to avoid pollution that harms the public. A high accuracy of the prediction model is required so that the predicted model can reach the true value. Forecasting models must be built on solid historical data to achieve high forecasting accuracy. However, data collected through various approaches are subject to ambiguity and uncertainty, resulting in less reliable predictive models. Therefore, the data must be handled accurately, to eliminate data uncertainty. Standard data processing processes are easy to use but do not provide a consistent method for dealing with this ambiguous data. Therefore, a method to deal with data containing uncertainty for forecasting purposes is presented in this paper. A new technique for providing uncertainty-based data preparation has been employed to develop an ARIMA-based model of environmental noise pollution. During the data preparation stage, the standard deviation approach was used. Prior to the development of the prediction model, it is crucial to manage the fuzzy data to minimize errors. The experimental findings show that the suggested data preparation strategy can increase the model's accuracy.

**Keywords**—Noise pollution; forecasting; ARIMA; uncertainty; standard deviation

## I. INTRODUCTION

One of the environmental challenges that has an impact on people's quality of life and well-being is noise pollution. Regular exposure to high levels of noise that might be damaging to people or other living beings is referred to as noise pollution [1] – [2]. Transportation noise, construction noise, manufacturing noise, and other extreme noise sources can all contribute to noise pollution [3] – [4]. Noise pollution has a long-term and short-term impact on human health, particularly hearing and mental health [5]. This causes awareness and action must be taken by the parties involved so that the effects of excessive noise do not cause emotional, mental, and physical health problems of people around the place involved.

Because diverse sounds occur on the construction site, the noise intensity might be extremely dangerous. Sound standard limits established by the Occupational Safety and Health Administration (OSHA) need ambient measurement. Noise levels at a construction site might vary based on the type of project and its stage of completion, whether indoors or outside

[6] – [7]. As the stage is completed, all activities on the building site change. This demonstrates that noise levels at a given stage are not consistent; they can be low or high [8] – [9]. Early stages of a construction site project, for example, include carpenters, cement workers, steelworkers, roofers, and bricklayers. Carpenters, ventilation installers, electricians, and plumbers begin their work in the following stages at the same time as drywallers, painters, and floor and ceiling installers. Depending on the task at hand, each of these stages employs a different set of tools. This will result in a wide range of noise, some of which is hazardous if it exceeds the standard noise limit [10].

Noise pollution has developed as a major environmental issue with substantial implications that are both stressful and harmful to one's health. Efforts must be made to reduce pollution. As with noise pollution, management must determine what steps should be taken to limit emissions. Forecasting is necessary for stakeholders to make better decisions and establish data-driven initiatives. It boosts management's confidence in making critical decisions. Because forecasting has become an important aspect of the planning process, particularly strategic planning, the establishment of a noise pollution forecasting model is critical [11]. To a large extent, the accuracy of management decisions is dependent on precise forecasting. Noise pollution projections are the foundation for efficient pollution management strategies as a preventive measure [12]. Much development of forecasting models has been done with noise pollution [13], [14], [15]. Noise prediction in construction sites [12], [16]. Statistical techniques, data analysis, and data mining are also used to model forecasts for noise pollution. Statistical techniques, data analysis, and data mining are also used to model forecasts for noise pollution. The findings of the study have supported this issue although improvements are needed to study other issues that arise.

Noise pollution data may involve uncertainty due to measurement error produces during pollution exposure assessment [17]. The measurement error is characterized by instrument imprecision and spatial variability [18]. Fault during measuring instruments in the technique used in the experiment give rise to uncertainty involvement in data collection [19]. Since most of the data comes from secondary sources, it could have problems with validity, bias, and representation. These problems could all lead to data inaccuracies and inefficient



forecasting models [20]. In short, the collected data or measured value contains uncertainty. When the data with uncertainty is analyzed to build a forecasting model, the uncertainty is carried through to the results, and thus reduces the model's accuracy. Handling uncertainty in data is one of the main challenges in forecasting. A widely used method to address the uncertainty lies in fuzzy theories [21]. Uncertainty involving noise pollution data has successfully used the fuzzy theory as a solution. Many studies have been produced with new solution model variants using fuzzy and hybrid concepts with other techniques [22] – [24].

Most studies in the literature focus on model development rather than data preparation. Good models, on the other hand, are often made from good data. Minimizing data collection and preparation errors can assist in the development of more accurate prediction models. Measuring the data collection accuracy is crucial to lower the chance of hidden errors in the created model. This paper proposes a systematic data preparation strategy for dealing with uncertainty during data preparation for forecasting, as measurement inaccuracy might have substantial implications for understanding noise pollution predictions. Time-series are used to store single-point data values and are best suited for classic time-series analysis techniques like autoregressive. The presence of underlying uncertainties, however, makes standard analysis ineffective in dealing with such data [25]. Considering this, addressing data uncertainties during data preparation is necessary as a stage in developing forecasting models. To solve the issues, this research introduces a new method of time series data preparation by modifying the spread of the symmetry triangular fuzzy number in the construction of autoregressive models.

In this paper, an experimental design for predicting ambient noise pollution by using ARIMA is established. In the data preparation stage, improvements were made due to the presence of fuzzy data. To obtain acceptable fuzzy values before building a prediction model, a systematic data preparation approach involving the translation of fuzzy data from a non-fuzzy number to a fuzzy number is required. The development of a triangular fuzzy number (TFN) that overcomes measurement uncertainties is offered as a systematic approach. To generate a triangular fuzzy number symmetry in the noise pollution data set, this strategy uses the standard deviation method to detect the triangle spread. Focus is placed on the preparation of fuzzy data and model forecasting, both of which are critical to explore to improve prediction value and accuracy.

The remainder of the paper is divided into the following subsections. Section 2 provides the theoretical foundations of this work, which include ARIMA and the triangular fuzzy number. Section 3 provides a description of the ARIMA-fuzzy solution approach. The proposed strategy is illustrated empirically in Section 4 using data on noise pollution. A few final observations are made in Section 5 to wrap things up.

## II. RELATED WORK

The ARIMA model has three parameters,  $p$ ,  $d$ ,  $q$  respectively corresponding to AR, I, MA. These three parameters affect the performance of ARIMA. In the

Autoregressive (AR) model, the current value of a variable is equated with the weighted sum of a set of part values and a completely random variation with the previous process and shock values. The  $p^{th}$  order autoregressive model AR( $p$ ), representing the variable  $y_t$  is generally written as follows:

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \dots + \phi_p y_{t-p} + e_t \quad (1)$$

where  $c$  is constant,  $e_t$  is white noise (error), and  $y_{t-1}$ ,  $y_{t-2}$ , ...,  $y_{t-p}$  are past series. The process at some point in time is the result variable in AR (1), and  $t$  is only related to line periods that are one period apart.

$I$  is the number of times the differential sequence must be repeated until it reaches a stationary state. For the ARMA model to work, it must be stationary.

Meanwhile, a common strategy for modeling univariate time series is the moving average (MA) process. The output variable of the moving-average model is defined as being linearly dependent on the current and various historical values of a stochastic (imperfectly predictable) factor. The  $q^{th}$  order autoregressive model MA( $q$ ), representing the variable  $y_t$  is generally written as Equation (2).

$$x_t = \mu + \phi_1 w_1 + \phi_2 w_2 + \dots + \phi_p w_p + w_t \quad (2)$$

where  $\mu$  is the mean of the series and  $w_t$ ,  $w_{t-1}$ ,  $w_{t-2}$ , ...,  $w_{t-q}$  are white noise (error).

### A. Triangular Fuzzy Number (TFN)

The fuzzy number has been introduced to deal with imprecise numerical quantities in a practical way [26] and has commonly been used by researchers [27].

Definition 4: Let  $a$ ,  $b$ , and  $c$  be real numbers with ( $a < b < c$ ). Then the Triangular Fuzzy Number (TFN)  $A = (a, b, c)$  with membership function is as follows:

$$y = m(x) = \begin{cases} \frac{x-a}{b-a}, & x \in [a, b] \\ \frac{c-x}{c-b} & x \in [b, c] \\ 0 & x < a \text{ and } x > c \end{cases} \quad (3)$$

We define TFN as Equation (4),

$$\tilde{y} = [\alpha_l, c, \alpha_r] \quad (4)$$

where  $c$  is the center,  $\alpha_l$  is the left spread, and  $\alpha_r$  is the right spread of the TFN. Symmetry TFN  $\tilde{y}$  has the same spread where  $c - \alpha_l = \alpha_r - c$ , and is denoted as:

$$\tilde{y} = [c, \alpha] \quad (5)$$

$\alpha$  is the spread of triangular fuzzy numbers. If  $\alpha = 0$ ,  $\tilde{y}$  is a non-fuzzy number.

The information provided is utilized to create ARIMA forecasting models with data that has been processed using fuzzy methods. The data used to build the ARIMA prediction model must first be pre-processed with triangular fuzzy numbers since it contains fuzzy elements. From previous literary works, fuzzy theory has been successfully applied widely in various case areas to reduce uncertainty and inaccuracy. However, few studies have clarified the details of fuzzy data preparation, which addresses data uncertainty.



Typically, expert definitions are used to define fuzzy numbers to address ambiguity. The definition of an expert, on the other hand, may be difficult to obtain and inconsistent, making it a difficult effort. Thus, in this study, a standard approach has been introduced to address uncertainties during data processing. The next section will describe the ARIMA prediction experiment that uses triangular fuzzy numbers as the data preparation method.

### III. FUZZY ARIMA MODELING

The process of cleaning and converting raw data prior to processing and analysis is known as data preparation. Reformatting data, making data adjustments, and integrating data sets to enrich data are all part of this crucial stage before processing. For data specialists or business users, data preparation may be time-consuming, but it is critical to place data in context to turn it into insights and reduce bias caused by poor data quality.

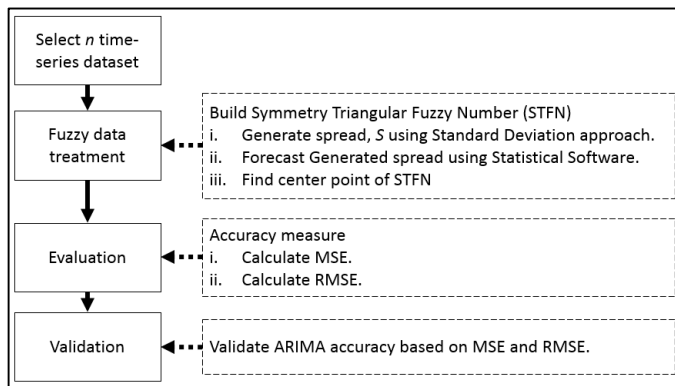


Fig. 1. Design of Experiment for ARIMA based Fuzzy Data Preparation.

The steps for building ARIMA with standard deviation-based symmetry triangular fuzzy number,  $\Delta_s$  is explained as follows:

Step 1. Select times series datasets as input data. Table I shows the input data format.

TABLE I. INPUT DATA FORMAT

Data	1	2	...	$n$
Input	$y_1$	$y_2$	...	$y_n$

Step 2. Construct a fuzzy number of symmetrical triangles with spreads generated by the standard deviation method,  $\Delta_s$ .

1) *Generate spread, S*: In Equation (6), a fuzzy time-series data  $\tilde{y}_t^s$  at a time,  $t$  is written with symmetry triangular fuzzy number data.

$$\tilde{y}_t^s = [y_t - s, y_t, y_t + s] \quad (6)$$

where  $y_t$  is time-series data at a time,  $t(t = 1, 2, \dots, n)$  and  $\Delta_s$  is the triangular spread based on the standard deviation of the dataset. The concept of standard deviation, which measures the spread of data, strikes at the core of this strategy.

2) Forecast generated spread using statistical software.

3) Find center value for symmetry triangular fuzzy number,  $\bar{y}_t^s$ .

The center value for the symmetry triangular fuzzy number is computed as follows:

$$\bar{y}_t^s = 0.5(\tilde{y}_t^s + \bar{y}_t^s) \quad (7)$$

where  $\bar{y}_t^s$  represent a center point for the triangle.  $\tilde{y}_t^s$  and  $\bar{y}_t^s$  represent left predicted value and right predicted value, respectively. Table II shows the data format of the center point for symmetry triangular fuzzy number,  $\bar{y}_t^s$ .

TABLE II. DATA FORMAT FOR CENTER POINT,  $\bar{y}_t^s$

$y_t$	$y_1$	$y_2$	...	$y_n$
$\bar{y}_t^s$	$\bar{y}_1^s$	$\bar{y}_2^s$	...	$\bar{y}_n^s$

Step 3. Calculate the Mean Squared Error (MSE).

The results are analyzed after all of the datasets have been tested. MSE is used to evaluate the accuracy's performance. The MSE for each  $y_t^s$  is calculated using Equation (8).

$$MSE = \sum_{i=1}^n \frac{(y_t - \bar{y}_t^s)^2}{n} \quad (8)$$

where  $y_t$  is a time-series data and  $\bar{y}_t^s$  are a predicted times series data at a time,  $t(t = 1, 2, \dots, n)$  and  $n$  is a sample size.

Step 4. Calculate the Root Mean Square Error (RMSE)

The RMSE is also calculated to help with the analysis. The RMSE for each  $y_t^s$  is calculated using Eq. 9.

$$RMSE = \sqrt{MSE} = \sqrt{\sum_{i=1}^n \frac{(y_t - \bar{y}_t^s)^2}{n}} \quad (9)$$

where  $y_t$  is a time-series data and  $\bar{y}_t^s$  are a predicted times series data at a time,  $t(t = 1, 2, \dots, n)$  and  $n$  is a sample size.

Step 5. MSE and RMSE values are used to validate ARIMA with  $\Delta_s$  accuracy. The model with the least MSE and RMSE has a higher prediction accuracy.

Model building refers to the process of deciding what model to use for the context. Sometimes, existing well-supported theory or knowledge guides the choice of model, but sometimes the choice needs to be made empirically, which is based on real data. Ordinary Least Squares (OLS) is a linear regression technique utilized in this study to infer the association between a variable and an outcome, especially when other factors are present. The coefficient and constant of linear regression are calculated using OLS and used to construct a linear regression model. When employing interval data, the method for obtaining the center point is critical for the validation phase.

The ARIMA model is built with fuzzy number production from a single point value to resolve uncertainties, as indicated in the systematic methods described in this section. Because crisp reliability is insufficient to grasp data-inbuilt uncertainties, this phase is critical during the data preparation process [28] – [29].

IV. EXPERIMENTAL RESULT AND DISCUSSION

Noise data collected at the construction site were used to evaluate the performance of the methods proposed in this section. The ARIMA model of this data set is ARIMA (1,0,0). The data set used has 1000 data points and covers the period from March 1, 2020, to March 31, 2020. Data was collected using a web-based Environmental Site Monitoring (IoT) System [30]. The technology is designed to monitor ambient noise throughout the day and provide real-time sound updates. Favoriot platforms collect and store data on cloud servers.

Step 1. Select noise time-series datasets. Table III shows noise datasets.

TABLE III. NOISE DATASETS

Data	1	2	...	999	1000
Noise	74	58.8	...	81.5	80

Step 2. Build symmetry triangular fuzzy number based on Standard Deviation method,  $\Delta_s$ .

1) Generate spread,  $S$ : The standard deviation approach (see Section 3 step 2) is used to calculate the spread of symmetry triangular fuzzy numbers, which is based on Eq (6).  $(\tilde{y}_t^s, y, \tilde{y}_t^s)$  represents a symmetrical triangle fuzzy number with a standard deviation-based spread. Table IV depicts the range of possible numbers of fuzzy symmetrical triangles.

TABLE IV. SYMMETRY TRIANGULAR FUZZY NUMBER SPREAD

$n$	$y_1$	$y_2$	...	$y_{999}$	$y_{1000}$
$\tilde{y}_t^s$	66.6956	51.4956	...	74.1956	72.6956
$\tilde{y}_t^s$	81.3044	66.1044	...	88.8044	87.3044

2) Forecast generated spread using statistical software. The predicted result for the spread shown in Table V.

TABLE V. PREDICTED RESULT FOR THE SPREAD OF SYMMETRY TRIANGULAR FUZZY NUMBER

$n$	$y_1$	$y_2$	$y_3$	...	$y_{23}$
$\tilde{y}_t^s$	-	12.7987	12.9442	...	14.6047
$\tilde{y}_t^s$	-	14.1413	14.2848	...	15.9473

3) Find the center value of symmetry triangular fuzzy number,  $\tilde{y}_t^s$  using Eq. (7). To calculate the MSE, the predicted value is transformed from the symmetry triangular fuzzy number to a single point. Table VI shows the symmetrical triangle fuzzy number's center value.

TABLE VI. CENTER POINT VALUE FOR THE SPREAD OF SYMMETRY TRIANGULAR FUZZY NUMBER

$y_t$	$y_1$	$y_2$	$y_3$	...	$y_{999}$	$y_{1000}$
$\tilde{y}_t^s$	-	74.5032	64.6617	...	77.6864	77.3949

Step 3. The MSE for the noise data is calculated based on Eq. (8), and then presented in Table VII.

TABLE VII. MSE FOR NOISE DATA

Data	ARIMA	ARIMA $\Delta_s$
Training	33.8190	19.6161
Testing	33.8194	17.0045

Step 4. The RMSE for the noise data is calculated based on Eq. (8), and then presented in Table VIII.

TABLE VIII. RMSE FOR NOISE DATA

Data	ARIMA	ARIMA $\Delta_s$
Training	5.8154	4.4401
Testing	5.8154	4.1339

Step 5. Validate ARIMA with  $\Delta_s$  accuracy based on MSE and RSME.

The MSE and RMSE results are compared to verify the correctness of the prediction error. This proposed method is also compared to traditional autoregressive (AR), autoregressive with standard deviation based (AR $\Delta_s$ ), and conventional autoregressive moving average methods (ARIMA). Table IX summarizes the MSEs for the noise data.

The MSE and RMSE results are compared to ensure that the prediction error is correct. Traditional autoregressive (AR), autoregressive with standard deviation based (AR $\Delta_s$ ), and conventional autoregressive moving average methods are also compared to this proposed method (ARIMA). The MSEs for the noise pollution data are shown in Table IX.

TABLE IX. SUMMARY OF MSE

Data	AR(1)	AR(1) $\Delta_s$	ARIMA	ARIMA $\Delta_s$
Training	52.3412	33.8197	*33.8190	33.8194
Testing	173.2402	18.1800	19.6161	**17.0045

\* Smallest MSE for Training  
\*\* Smallest MSE for Testing

The results in Table IX show good enforcement when compared to the typical strategy. The proposed technique can achieve higher accuracy than traditional AR and conventional ARIMA. In the AR model, the proposed technique drove the MSE from 173.2402 to 18.10, and in the ARIMA model, it pushed the MSE from 19.6161 to 17.0045. To improve the outcome, the RMSE approach was also used. The RMSEs for the noise pollution data are summarized in Table X.

TABLE X. SUMMARY OF RMSE

Data	AR(1)	AR(1) $\Delta_s$	ARIMA	ARIMA $\Delta_s$
Training	7.2347	5.8155	*5.8154	*5.8154
Testing	13.1949	4.2744	4.4401	**4.1339

\* Smallest MSE for Training  
\*\* Smallest MSE for Testing

The outcomes in Table X surpass those in Table IX and are consistent with the standard model. The AR model's MSE may be increased from 13.1949 to 4.2744, and the ARIMA model's MSE can be increased from 4.4401 to 4.1339 using this strategy. The MSE and RMSE decrease as the results improve.

It has been demonstrated that symmetric triangular fuzzy numbers can be created using standard deviations. As shown in Tables IX and X, MSE and RMSE for ARIMA with standard deviation based are superior to other techniques. Eqs (9) and (10) show the prediction model.

$$AR(1) = ARIMA(1) = 26.59 + 0.70y_{t-1} \quad (9)$$

$$AR(1)_{\Delta_s} = ARIMA(1)_{\Delta_s} = (24.02, 29.17) + 0.65y_{t-1} \quad (10)$$

#### V. CONCLUSION

Data preparation is critical and is a necessary step before developing forecasting models. Hence, data processing is required, and it must be carried out using proper procedures to manage data errors. Furthermore, data preparation is critical for producing high-quality data. This is done by organising and reformatting the data set and ensuring the high quality of the data used in the study. A strong forecasting model can only be built with high-quality input data; hence this is an essential prerequisite. In addition, the provision of data may help relevant parties to make better business decisions. This is because fast, effective and high-quality business decisions are produced when high-quality data is handled, examined and processed more quickly and efficiently.

To address the uncertainties in the data, we describe a technique for creating symmetric triangular fuzzy integers with standard deviation. It offers a simple and easy-to-implement solution. This experiment compares the results of the suggested approach using a few different methods. The effectiveness of the suggested method has been described and evaluated against the current method. The experimental results demonstrate that this proposed strategy using symmetric triangular fuzzy numbers outperforms the others in terms of predicted accuracy. In short, symmetric triangular fuzzy numbers are one of the other approaches to improve time series forecasting outcomes, particularly in this experiment with noise pollution.

#### ACKNOWLEDGMENT

This research was supported by the Ministry of Education (MOE) through the Fundamental Research Grant Scheme (FRGS/1/2019/ICT02/UTHM/02/7) Vot K208. This research work is also supported by the Ministry of Education, R.O.C., under the grants of TEEP@AsiaPlus. The work of this paper is also supported by the Ministry of Science and Technology under Grant No. MOST 109-2221-E-035-063-MY2.

#### REFERENCES

- [1] Mohamed, A. M. O., Paleologos, E. K., & Howari, F. M. "Noise pollution and its impact on human health and the environment," In *Pollution assessment for sustainable practices in applied sciences and engineering*, 2021, (pp. 975-1026). Butterworth-Heinemann.
- [2] Morillas, J. M. B., Gozalo, G. R., González, D. M., Moraga, P. A., & Vilchez-Gómez, R. "Noise pollution and urban planning," *Current Pollution Reports*, vol. 4, no. 3, pp. 208-219, 2018.
- [3] Patel, D. B., & Solanki, H. K. A. "Effects of Noise Pollution on Human Health," *Research and Reviews: Journal of Environmental Sciences*, vol. 3, no. 1, pp. 1-5, 2021.
- [4] Farooqi, Z. U. R., Sabir, M., Latif, J., Aslam, Z., Ahmad, H. R., Ahmad, I., ... & Ilić, P. "Assessment of noise pollution and its effects on human health in industrial hub of Pakistan," *Environmental Science and Pollution Research*, vol. 27, no. 3, pp. 2819-2828, 2020.
- [5] Hasan, S. W., & Jamal, M. A. "Noise Pollution, Effect of Noise on Behaviour of Animals and Human Health," *Texas Journal of Medical Science*, vol. 1, no. 1, pp. 71-75, 2021.
- [6] Ning, X., Qi, J., Wu, C., & Wang, W. "Reducing noise pollution by planning construction site layout via a multi-objective optimization model," *Journal of cleaner production*, vol. 222, pp. 218-230, 2019.
- [7] Subramaniam, M., Hassan, M. Z., Sadali, M. F., Ibrahim, I., Daud, M. Y., Aziz, S. A., ... & Sarip, S. "Evaluation and analysis of noise pollution in the manufacturing industry," In *Journal of Physics: Conference Series*, vol. 1150, no. 1, pp. 012019, 2019.
- [8] Feng, C. Y., Noh, N. I. F. M., and Al Mansob, R. "Study on The Factors and Effects of Noise Pollution at Construction Site in Klang Valley," *Journal of Advanced Research in Applied Sciences and Engineering Technology*, 20(1), pp. 18-26, 2020.
- [9] Mir, M., Nasirzadeh, F., Lee, S., Cabrera, D., & Mills, A. "Construction noise management: A systematic review and directions for future research," *Applied Acoustics*, vol. 197, pp. 108936, 2022.
- [10] OSHA Region III. (2001). "Enforcement of the Occupational Noise Exposure Standards," 29 CFR 1910.95, 1926.52, and 1926.101, Inspection Procedures and Interpretive Guidance.
- [11] Mohammed, M. U., Badamasi, M. M., Usman, F., Zango, Z. U., Dennis, J. O., Aljameel, A. A. I., ... & Hussein, T. M. "Towards Urban Sustainability: Developing Noise Prediction Model in an Informal Setting," *Applied Sciences*, vol. 12, no. 18, pp. 9071, 2022.
- [12] Kwon, N., Lee, J., Park, M., Yoon, I., and Ahn, Y. "Performance evaluation of distance measurement methods for construction noise prediction using case-based reasoning," *Sustainability*, vol. 11, no. 3, pp. 871, 2019.
- [13] Lu, H., & Wang, T. "An Automobile Noise Prediction Model Based on Extension Data Mining Algorithm," *Rev. d'Intelligence Artif.*, vol. 33, no. 5, pp. 341-347, 2019.
- [14] Sharma, A., Vijay, R., Bodhe, G. L., and Malik, L. G. "An adaptive neuro-fuzzy interface system model for traffic classification and noise prediction," *Soft Computing*, vol. 22, no. 6, pp. 1891-1902, 2018.
- [15] Awan, F. M., Minerva, R., and Crespi, N. "Using Noise Pollution Data for Traffic Prediction in Smart Cities: Experiments Based on LSTM Recurrent Neural Networks," *IEEE Sensors Journal*, vol. 21, no. 18, pp. 20722-20729, 2021.
- [16] Buxton, R. T., McKenna, M. F., Mennitt, D., Fristrup, K., Crooks, K., Angeloni, L., and Wittemyer, G. "Noise pollution is pervasive in US protected areas," *Science*, vol. 356, no. 6337, pp. 531-533, 2017.
- [17] Nemes, A., Mester, G., and Mester, T. "A Soft Computing Method for Efficient Modelling of Smart Cities Noise Pollution," *Interdisciplinary Description of Complex Systems: INDECS*, vol. 16, no. 3-A, pp. 302-312, 2018.
- [18] Murphy, E., Faulkner, J. P., & Douglas, O. "Current state-of-the-art and new directions in strategic environmental noise mapping," *Current pollution reports*, vol. 6, no. 2, pp. 54-64, 2020.
- [19] Lagonigro, R., Martori, J. C., & Apparicio, P. "Environmental noise inequity in the city of Barcelona. Transportation Research Part D: Transport and Environment," vol. 63, pp. 309-319, 2018.
- [20] Ghaderi, M., Javadikia, H., Naderloo, L., Mostafaei, M., and Rabbani, H. (2019). "Analysis of noise pollution emitted by stationary MF285 tractor using different mixtures of biodiesel, bioethanol, and diesel through artificial intelligence," *Environmental Science and Pollution Research*, vol. 26, no. 21, pp. 21682-21692, 2019.
- [21] L. A. Zadeh, "Fuzzy Sets," *Inf. Control*, no. 8, pp. 338-353, 1965.
- [22] De, S. K., Swain, B. K., Goswami, S., and Das, M. "Adaptive noise risk modelling: fuzzy logic approach," *Systems Science & Control Engineering*, vol. 5, no.1, pp. 129-141, 2017.
- [23] Saleh, A., Rosnelly, R., Puspita, K., and Sanjaya, A. (2017, August). "A comparison of Mamdani and Sugeno method," for optimization prediction of traffic noise levels. In *2017 5th International Conference on Cyber and IT Service Management (CITSM)* (pp. 1-4). IEEE.
- [24] Singh, D., Upadhyay, R., Pannu, H. S., and Leray, D. "Development of an adaptive neuro fuzzy inference system based vehicular traffic noise prediction model," *Journal of Ambient Intelligence and Humanized Computing*, vol. 12, no. 2, pp. 2685-2701, 2021.

- [25] Bray, S., Caggiani, L., and Ottomanelli, M. "Measuring transport systems efficiency under uncertainty by fuzzy sets theory-based Data Envelopment Analysis: theoretical and practical comparison with traditional DEA model," *Transportation Research Procedia*, 5, pp. 186-200, 2015.
- [26] Pisz, I., Chwastyk, A., and Łapuńka, I. "Assessing the profitability of investment projects using ordered fuzzy numbers," *LogForum*, vol. 15, no. 3, pp. 377-389, 2019.
- [27] Triangular Fuzzy Numbers.
- [28] Lah, M. S. C., Arbaiy, N., and Lin, P. C. "Stock Index Modelling Using Arima with Standard Deviation Based Triangular Fuzzy Numbers," *Journal of Critical Reviews*, vol. 7, no. 8, pp. 1264-1268, 2020.
- [29] Lin, P. C., Lah, M. S. C., and Arbaiy, N. "An Architecture of Air Pollution Alert System based on ARIMA Model," *Solid State Technology*, 63(3), pp. 881-892, 2020.
- [30] Arbaiy, N., Sapuan, S. A., Lah, M. S. C., Othman, M. H. H., and Lin, P. C. "The Construction Site Ambient Noise Monitoring System with Internet of Things (IoT)," *Comput. Res. Prog. Appl. Sci. Eng. CRPASE*, vol. 5, no. 4, pp. 118-121, 2019.

# Extractive Multi-document Text Summarization Leveraging Hybrid Semantic Similarity Measures

Rajesh Bandaru, Dr. Y. Radhika

Department of CSE, GST, GITAM (Deemed to be University)  
Visakhapatnam, India

**Abstract**—Because of the massive amount of textual information accessible today, automated extraction text summarization is one of the most extensively used ways to organize the information. The summarization mechanisms help to extract the important topics of data from a given set of documents. Extractive summarization is one method for providing a representative summary of a text by choosing the most pertinent sentences from the original text. Extractive multi-document text summarization systems' primary goal is to decrease the quantity of textual information in a document collection by concentrating on the most crucial subjects and removing irrelevant material. In the previous research, there are several methods such as term-weighting schemes and similarity metrics used for constructing an automated summary system. There are few studies that look at the performance of combining various Semantic similarity and word weighting techniques in automatic text summarization. We evaluated numerous semantic similarity metrics in extractive multi-document text summarization in this research. In the extractive multi-document text summarization discussed in this research, we looked at numerous semantic similarity metrics. ROUGE metrics have been used to evaluate the model performance in experiments using DUC datasets. Even more, the combination formed by different semantic similarity measures obtained the highest results in comparison with the other models.

**Keywords**—*Extractive text summarization; semantic similarity; sentence scoring; summary*

## I. INTRODUCTION

The amount of data and information available has exploded since the introduction of the World Wide Web. The volume of data has grown to the point where it is nearly difficult for any specific firm to analyze it all, or to summarize it. People are reluctant to engage in reading a lengthy piece of text and, as a result, typically skip crucial sections of it. This has boosted the need for text summarization automation [1].

In general, a person follows the three processes outlined below to create a summary: 1) interpreting the document's content, 2) selecting relevant chunks of meaningful information, 3) putting this content of data. Because of their difficulties, there is limited possibility of automating the first and third processes for any random text. As a result, the majority of techniques aim to automate the second phase [1].

Text summarizing is considered as single or multi-document summary terms of the number of documents studied and summarized at the same time. In single document summarizing, a summary is constructed from a single

document, but in multi-document summarization, a series of documents is examined for creating a summary. The task of summarizing many papers is more complex than the process of summarizing a single document. One of the most difficult issues in summarizing many publications is redundancy. Additional classification categories, such as single vs. multi-document categorization and mono-lingual vs. multi-lingual summarization, have been developed in the past based on many other factors [2].

Text summarization techniques often are divided into two categories: abstractive and extractive. The primary goal in extractive summarization included to retrieve the most essential sentences from a document(s) and combine them into a summary. This is in contrast to abstractive summarization, which involves reiterating the information in the text. The extractive summary contains sentences taken directly from the original content, whereas an abstract summary uses terms / expressions not present in the original source. Because of its greater practicality, extractive summarization has become a benchmark in text summarizing [3]. Abstractive text summarizing techniques try to generate summaries that summarize the crux of the text in the same way as people do after studying any text. This employs generative methodologies that can produce meaningful phrases while maintaining the semantics of the source text. This is regarded as a tough topic to tackle, and several novel ways have been presented.

Extractive summarization is divided into three stages: pre-processing, sentence scoring, and sentence selection. Several activities, like as tokenization, phrase and paragraph segmentation, are often carried out during the pre-processing phase. During the sentence scoring step, sentences are ranked based on certain criteria, and every sentence is assigned a score. Finally, the finest sentences are chosen and incorporated in summary during the sentence selection process. As previously stated, one of its most significant issues in multi-document summarizing is duplication, because identical phrases are more likely to be encountered in distinct documents, frequently [4].

Extractive text summarization is a simpler and more reliable method of creating summaries in which key lines from a text are chosen and provided to the user. Each sentence is scored, and the sentences with the highest scores are chosen for inclusion in the extract [5]. This is significantly easier than abstractive summaries, which need the production of phrases and words, as well as their organization into legible sentences, while yet giving an understandable substance of the subject. It

would need a significant amount of natural language processing, making it a significantly more complex task.

With the use of data-driven methods and semantic similarity approaches, extractive summaries will be produced in this study in order to meet the goal of text summarization. This involves analysing the large volume of information and creating a list of the sentences that could be the most helpful and contain the main idea of the text. Although most of the time, people strive to summarize texts in a way that conveys the same sense as the original text and do not see summaries as phrases taken literally from the source [6].

The remainder of the paper can be found in the sections that follow this one: Section 2 presents the results of a survey of the literature on text summarization using various approaches, which was carried out in order to create this paper. After providing a thorough explanation of the proposed algorithm in Section 3, Section 4 presents the results of the experiments conducted as part of the research. After a discussion of the findings and recommendations for future research are in Section 5.

## II. RELATED WORK

Since the 1950s, researchers have been studying automatic text summarization. It has since been extensively researched. Researchers working on document summarizing all around the globe are experimenting with a variety of approaches in order to produce ways that deliver the highest results. This work focuses on extractive multi-document summarization.

Various extraction-based strategies for generic multi-document summarization have been suggested so far. Statistical techniques deal with statistical aspects that aid in the extraction of relevant phrases and words from source material. Furthermore, traits and their weights play a significant influence in establishing sentence relevance. This section presents various models employed in the domain of multi-document summarization.

Jesus M. Sanchez-Gomez et al. [5] proposed a model with a set of multi objective functions. The objective functions defined in this work targets to coverage of content and reducing the redundancy. Using a combination of statistical and graph-based methods Mohammad Bidoki et al. [6] proposed a semantic framework for developing an extractive multi-document summarizer system. It is a dialect, unsupervised system. To learn the semantic representation of words from a set of supplied documents, the model uses the word2vec technique. It expands on each phrase using a one-of-a-kind method that employs the most informative and least repetitive words related to the statement's fundamental idea. Phrase expansion implicitly achieves word meaning disambiguation and adapts conceptual density to each sentence's main idea. The importance of sentences is then determined using the graph representation of the documents.

Begum Mutlu and colleagues [7] have created an English dataset including the proceedings of SIGIR 2018. Three readers used a manual labelling approach to classify the assertions in the opening parts as summary-worthy or summary-unworthy. It was shown that employing ensembled feature space considerably improved summarization

performance when both conventional classification and ROUGE-based analysis were used.

Hiren Kumar Thakkar et al [8] proposed a novel Domain Feature Miner (DOFM) mining algorithm. The summaries generated by DOFM are then subjected to automatic examination using ROUGE. This is a well-known programme for automated assessment of summaries. An error study revealed that 84 percent of the sentences from all DOFM generated summaries were selected by at least one of the three annotators. This highlights the DOFM's resiliency in terms of domain feature retrieval and extractive summarization, as well as its overall performance.

Ángel Hernández-Castañeda et al [9] uses Genetic Algorithm to identify the most effective grouping of words. This model organizes sentences in a text with the assistance of a clustering approach. Summaries generated not only contain matches of unique words, but also give context by matching terms in the text. One-of-a-kind technique for automated summarization is presented in this study. This method can be used to organize sentences in a text according to specific semantic and lexical qualities. It combines a vectorial space formed by a large number of feature generation algorithm(s) with a single summary strategy. It is not necessary to have a prior grasp of the underlying issue in order to generate vectors for this purpose. LDA, Doc2Vec, TF-IDF, and OHE do not need any prior knowledge of classes.

Kaichun Yao et al. [10] developed a extractive document summarizing approach based on Deep Q-Networks (DQN) to capture word salience and redundancy and train a strategy that maximises the Rouge score over gold summaries. The information given by the informative features not only provides informative features to describe the DQN's states but also generates a list of probable DQN actions from the document's words. Our model does not need extractive labels at the sentence level since it is trained directly on human-provided reference summaries. The Rouge measure is used to assess the model's performance on the CNN/Daily, DUC 2002, and DUC 2004 datasets. When applied to non-linguistic corpora, our technique outperforms or is on par with state-of-the-art models in terms of performance. The researchers believe this is the first time DQN has been used for extractive summarization in any scientific setting.

Luca Cagliero et al. [11] mention that annotating scientific articles with textual highlights, it is feasible to provide readers with potentially valuable result-oriented insights that may be used immediately. Unfortunately, the majority of the time, rather than automatically, the annotation process is performed by hand. A further problem is that the highlight information is completely lacking from the vast majority of earlier publications. The solution provided here overcomes the issues noted above by using supervised learning on previously annotated article data.

Jesus M. Sanchez-Gomez et al. [12] using three distinct term-weighting algorithms performed the task of multi-document text summarizing, and the authors found that they were all successful. Different unique similarity metrics that are employed in text-similarity have been taken into consideration by this work. The average and Pearson's coefficient of



variation are two of the computations that were employed in this investigation.

Mohammad Mojri et al. [13] proposed the MTSQIGA approach, which is a novel multi-document text summarization approach. It is designed to extract salient sentences from a source document collection in order to generate a summary of the information contained in the collection. It is proposed that a modified quantum measurement, as well as a self-adaptive quantum rotation gate, be used in conjunction with a summary generator that is dependent on the quality and length of the summary that is generated. A benchmark dataset from the DUC 2005 and 2007 was used to evaluate the proposed system in terms of ROUGE standard measures.

Akanksha Joshi et al. [14] proposed SummCoder, a method for extracting text summarization from single documents, makes this task much easier. The summary is generated using three metrics: content relevance, novelty, and position relevance. The following are the outcomes: An auto-encoder network is used to determine the relevance of sentence content by exploiting the similarity between embeddings in distributed semantic space, and the novelty metric is derived from this similarity. In this feature, which was created by hand, a dynamic weight calculation function based on the overall length of the document is used to give more weight to the document's first few sentences. It is also possible to create a document summary by ranking the sentences based on a combined final score derived from three different sentence selection metrics. A new summarization benchmark, the Tor Illegal Documents Summarization (TIDSumm) dataset, will benefit law enforcement agencies (LEAs). These summaries were created manually for 100 documents from onion websites in the Tor (The Onion Router) network and are included in the dataset. When compared to other methods, this text summarization approach achieves comparable or better performance for a wide range of ROUGE metrics for the DUC 2002, Blog Summarization, and TIDSumm datasets.

Manh et al. [30] used corpus based measures like LSA and LDA along with K-means to perform the task of summarization. The results of this work are comparatively good as corpus based measures explores different possibilities of evaluation. But the semantic similarity is not explored by Manh et al. [30]. The authors [31], [32] provided the applications of semantic similarity measures for the evaluating verb similarity and sentence with contradictory similarity. The works [31], [32] also discussed the importance of semantic similarity in the current research domains of natural language processing.

Table I provides the summary of the models compared in this article. Table I gives an insight into the models target and whether sentence scoring is performed in the model or not. This section presented various multi-document summarization models based on the different techniques. The models presented in this work are limited to analysing the similarity between the sentence and document title using term weighting schemes. The significance of knowledge based measures is not considered in the works highlighted in this section. We propose a novel knowledge based metric based on information content and path length in the next section.

TABLE I. SUMMARY OF THE RELATED WORK

Ref. No	Model	Target	Sentence Scoring
[5]	Artificial Bee Colony algorithm	Text summarization of multiple documents	Not mentioned
[6]	Semantic Approach	Text summarization of multiple documents	Mentioned
[7]	Candidate sentence selection	Text summarization of multiple documents	Mentioned
[8]	Domain Feature Miner	Text summarization of multiple documents	Not mentioned
[9]	Language-independent Summarization	Keyword Extraction enabled Text summarization	Not mentioned
[10]	Deep reinforcement learning	Text summarization	Not mentioned
[11]	Unsupervised framework	Auto encoder based Text summarization	Not mentioned
[12]	Supervised summarization	Scientific Article Summarization	Not mentioned
[13]	Centroid approach and sentence embeddings	Extractive Text Summarization	Mentioned
[14]	Term-weighting schemes	Text summarization	Mentioned
[15]	Quantum-inspired genetic algorithm:	Text summarization	Mentioned
[16]	Entropy for Extractive Document Summarization	Different measures of text summarization	Not mentioned
[17]	Weighted Word Embedding	Text summarization	mentioned
[18]	Firefly algorithm	Text summarization of multiple documents	Not mentioned
[19]	Fuzzy and evolutionary based model	Extractive Text Summarization	Mentioned
[20]	Summarization of documents	Summarization of image based documents	Not mentioned

The next section gives the proposed model and the significance of knowledge based measures in the sentence similarity evaluation.

### III. PROPOSED MODEL

This section presents the proposed work to perform the document summarization. This section also proposes a novel metric using the concepts of semantic similarity to estimate the values of sentence scoring. The first part of this section covers the knowledge-based measures and the second part covers the document summarization aspects.

The measures that produce synonyms and also deal with various word forms are knowledge-based measures. Similarity based on knowledge is determined by the information content or the length between the terms [22]. To determine similarity, knowledge-based methods make use of a well-constructed taxonomy (also known as a lexical database) to infer semantic similarity between concepts.

Information content (IC) is the probability regarding the availability of a concept in the corpus.

$$IC(C_k) = -\log(P(C_k)) \quad (1)$$

Where  $P(C_k)$ , is probability of the concept  $C_k$  in the corpus.

$$P(C) = \frac{f(C)}{N} \quad (2)$$

Where  $f(C)$  is the frequency of the concept in the corpus,  $N$  represents the number of words in the corpus.

Path (or path length) between two concepts gives the possible shortest path between the concepts.

The depth (D) is referred as the length of concepts and maximum depth (Dmax) of a concept the length from the concept to root in the taxonomy.

Least common subsumer (LCS) of two concepts in the taxonomy is another concept which is the root of the two concepts.

#### A. Semantic Similarity Measures

Various semantic similarity measures which are knowledge based are discussed in this subsection. The following are the standard knowledge based semantic measures.

Res Measure [21]: This measure estimates the similarity between the concepts  $C_i, C_j$  by considering the information content of the lowest common subsumer.

$$resnik(C_i, C_j) = IC(C_{LCS}(C_i, C_j)) \quad (3)$$

Jcn [22] Measure: This measure to calculate the similarity between the concepts  $C_i, C_j$  proposes the following equation,

$$jcn(C_i, C_j) = IC(C_i) + IC(C_j) - 2 * IC(C_{LCS}(C_i, C_j)) \quad (4)$$

Lin Measure [23]: This measure is defined as,

$$lin(C_i, C_j) = \frac{2 * IC(C_{LCS}(C_i, C_j))}{IC(C_i) + IC(C_j)} \quad (5)$$

Lch [24] Measure: This measure considers the maximum depth of the taxonomy and the length of concepts  $C_i, C_j$  in the taxonomy.

$$lch(C_i, C_j) = -\log\left(\frac{len(C_i, C_j)}{2 * D_{max}}\right) \quad (6)$$

Wup [25] (wup) Measure: This measure uses depth of lowest common subsumer of the two concepts  $C_i, C_j$  and individual depths of the concepts to estimate the similarity between concepts.

$$wup(C_i, C_j) = \frac{2 * D(C_{LCS})}{D(C_i) + D(C_j)} \quad (7)$$

Path Measure [26] (path): This measure calculates the inverse of semantic distance between the concepts  $C_i, C_j$  as the similarity between the concepts.

$$pathdistance(C_i, C_j) = length(C_i, C_j) \quad (8)$$

$$path(C_i, C_j) = \frac{1}{1 + pathdistance(C_i, C_j)} \quad (9)$$

Li measure [27] (li): This is a non-linear measure to estimate the similarity of the concepts. This measure uses depth and length between the concepts  $C_i, C_j$  to calculate the similarity.

$$li(C_i, C_j) = e^{-\alpha * l} * \frac{e^{\beta * D} - e^{-\beta * D}}{e^{\beta * D} + e^{-\beta * D}} \quad (10)$$

Where  $\alpha, \beta$  are parameters and  $\alpha = 0.2, \beta = 0.6$ .

#### B. Proposed Measure

The metrics indicate that they are all attempting to calculate how much information is shared between them in order to determine how closely two concepts are related. The problem of concepts with the same length and giving the same value even when there is less similarity between the concepts is also not addressed by measures based on the distance between concepts. The issue of equal route length can be solved by adding depth using techniques like wup and li. Greater granularity has the unintended consequence of making the concepts at the top of the hierarchy less detailed. This indicates that the path and depth problems are being addressed using the data. Compared to the route- and depth-based assessments, the information content measurements are more accurate. These measures will give the same similarity score for two concepts that have the same LCS, regardless of how differently their contents are expressed. For this problem, the information content serves as a guiding weight, and the measure may be represented as follows:

$$\text{Hybrid measure}(C_i, C_j) = \frac{1}{1 + path(C_i, C_j) * k^{IC(2 * C_{LCS}(C_i, C_j)) / (IC(C_i) + IC(C_j))}} \quad (11)$$

The suggested metric has various weights for the path length, which eliminates the issue of ideas having the same path length and hence having the same LCS difficulties. Conceptual similarity is estimated by taking into consideration the semantic distance between ideas and their respective information content as well as the information content of each concept measured separately (LCS).

The sentence scoring is calculated by deriving a sentence feature vector. The sentence feature vector is calculated by using the NLTK and proposed semantic similarity measure.

#### C. Proposed Extractive Multi-Document Summarization

In this section, we discuss our proposed system. Fig. 1 gives the architecture of the proposed model. The input to the model is a set of documents. The documents are taken from reliable datasets.

The first phase in the model is preprocessing of data. In the preprocessing, sentence segmentation is performed initially. Later the sentences are tokenized, and each sentence is represented as a set of tokens at this step. The parts-of-speech tagging relative to the words in the sentence is also preserved. The most irrelevant words from the sentence are removed and the remaining words are stemmed.

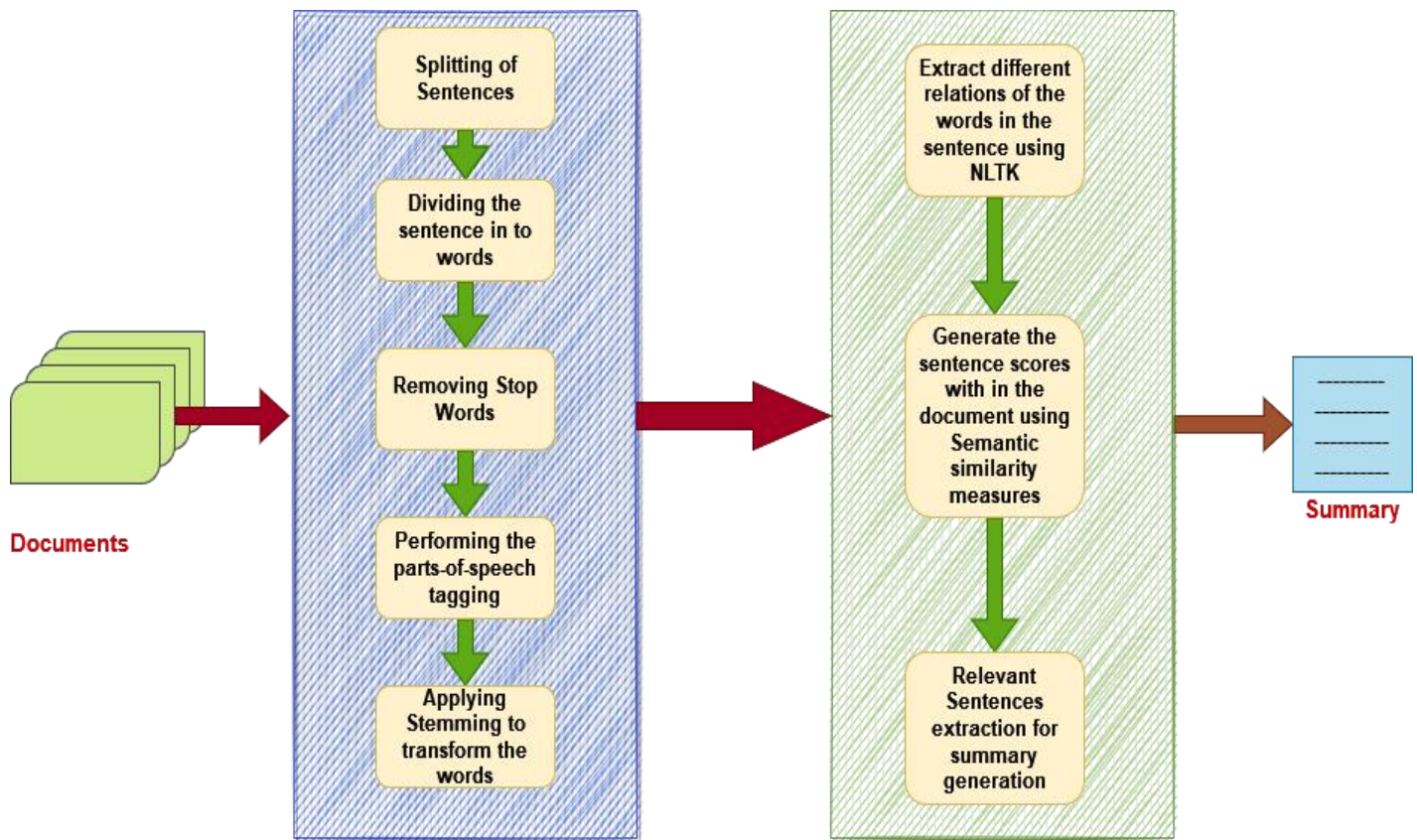


Fig. 1. Architecture of the Proposed Model.

The next stage in the proposed model is to extract multiple senses of each word preserved in the sentence. The relations are extracted from the NLTK package. The next step in the second stage is to generate the sentence scores. The sentence scores are used to generate the summary of the documents. Later the performance metrics are used in the evaluation.

Preprocessing is accomplished by the use of a pipeline that is often used for multi-document summarizing jobs, in which a cluster of documents is represented by a collection of phrases. In other words, we're discussing the presence of a cluster  $D$  containing  $m$  documents  $D = [d_1, d_2, \dots, d_m]$ .

The initial step was to decompose each document  $d_i$  in the cluster  $D$  into individual phrases, which we performed with the help of the free and open-source software package spaCy for Advanced Natural Language Processing. The next phase uses the Natural Language Toolkit (NLTK) and regular expressions to clean up these phrases by converting all words to lower case and deleting special characters, unnecessary whitespace, HTML elements, URLs, and email addresses from the source code.

#### D. Semantic Relationship between Words

We employ semantic similarity metrics and WordNet to capture the semantic links between words. We begin by assessing the word's resemblance to the document title. The highest degree of similarity that a word achieves is referred to as the word score. The TF-IDF, a term weighting approach, is used to weight these word scores.

We use multiplication to integrate TF-IDF and word scores in our current work. To create a sentence vector, we integrate all of the sentence's word weighed scores. The TF-IDF assists in mapping the phrase to a distributed semantic vector, with the exception that the most frequently occurring words have a smaller influence on the outcome. Finally, we acquire a vector representing all sentences in the corpus; this vector is referred to as the average sentence vector.

#### E. Extracting Different Features

The linguistic features of the sentence are also extracted as important features to calculate the sentence score. Representation of the calculations of the sentence vector using different features is mentioned in Fig. 2.

1) *Noun and verb phrase*: The noun or verb phrases in a sentence are essential and given more weight in a sentence. Each sentence's noun and verb phrase weight is computed as follows:

$$NV_{Phrases} = \frac{No. (Verbs, Nouns)}{Sentence Length}$$

2) *Sentence position*: In general, the sentences at the beginning and end are more informative.

3) *Sentence length*: Sentences with large length are more significant in the document.

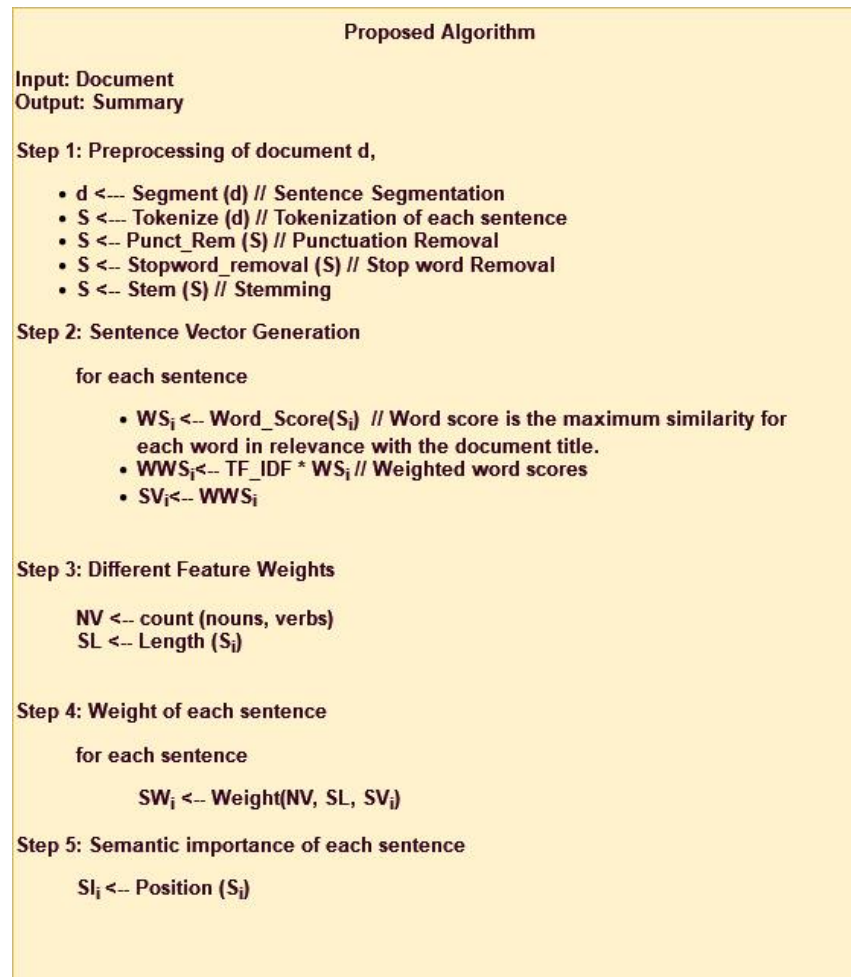


Fig. 2. Proposed Algorithm.

Based on the features and sentence vectors the sentence scores are generated and the sentences with higher scores are considered as more relevant sentences and these sentences are combined together to generate the summary of sentences. The different steps involved in calculating the sentence scores are mentioned in the proposed algorithm. The different features used for scoring the sentence are discussed above.

This section covers the overall description of the proposed model and the results regarding these models are presented in the next section.

#### IV. RESULTS AND DISCUSSION

This section covers the experimentation of the proposed similarity measure on word pair similarity dataset and the proposed model to perform multi-document summarization of data. The first part of the results is with respect to the semantic similarity on different word pair datasets.

##### A. Metrics Used

**Pearson correlation:** This correlation is used to evaluate the performance of various semantic similarity measures.

**Spearman Correlation:** This is also another well-known correlation that is used to evaluate the performance of various semantic similarity measures.

**ROUGE Score:** Many researchers and practitioners use this metric to assess how well multi-document summarizing algorithms function.

##### B. Datasets

- RG dataset [28]: This dataset, which includes 65 noun pairs, is used to assess word similarity tasks.
- MC dataset [29]: This dataset, which includes 30 noun pairs, is used to assess word similarity tasks.
- DUC 2007 dataset: The dataset consists of 45 separate subjects that are each covered by 45 unique documents, all of which cover all 45 categories.

##### C. Tools used for Implementation

- NLTK
- Spacy
- ROUGE
- SCIKIT learn
- Anaconda



The results of several semantic similarity tests performed on the RG dataset and the MC dataset are shown in Tables II and III, respectively. The findings show that combining the length between the concepts with the information content leads in greater correlation values. The proposed hybrid measure is able to achieve better results when compared with all the existing models.

As a baseline for our current research, we utilise the primary task dataset from the Document Understanding Conference (DUC 2007). Automated text summarization assessment is carried out by NIST using the DUC 2007 dataset. The DUC 2007 dataset is made up of news stories from a variety of publications. The dataset contains 45 separate subjects and 45 individual texts, each of which discusses all 45 themes.

Fig. 3 is an example of how summarization works. The proposed model after sentence scoring extracts the relevant sentences according to the document in the shown example. The results of the summarization are with respect to a compression rate of 15%. It can be observed from the figure that the relevant sentences are extracted according to the given data.

TABLE II. RESULTS OF CORRELATION ON RG DATASET

Measure	Spearman	Pearson
path	0.78	0.78
li	0.79	0.86
lin	0.78	0.86
res	0.78	0.84
lch	0.78	0.84
wup	0.76	0.79
jcn	0.78	0.72
<b>Proposed</b>	<b>0.80</b>	<b>0.86</b>

TABLE III. RESULTS OF CORRELATION ON MC DATASET

Measure	Spearman	Pearson
path	0.78	0.78
li	0.79	0.86
lin	0.78	0.86
res	0.78	0.84
lch	0.78	0.84
wup	0.76	0.79
jcn	0.78	0.72
<b>Proposed</b>	<b>0.80</b>	<b>0.86</b>

Because of the massive amount of textual information accessible today, automated extraction text summarization is one of the most extensively used ways to organise the information. The summarization mechanisms helps to extract the important topics of data from a given set of documents. One of the summarization techniques named extractive summarization selects the most relevant sentences from the text to provide a representative summary. The main purpose of extractive multi-document text summarization systems is to reduce the amount of textual information in a document collection by focusing on the most important topics and avoiding unnecessary portions. In the previous research, there are several methods such as term-weighting schemes and similarity metrics used for constructing an automated summary system. There are few studies that look at the performance of combining various Semantic similarity and word weighting techniques in automatic text summarization. We evaluated numerous semantic similarity metrics in extractive multi-document text summarization in this research. In this paper, we explored various semantic similarity measures in the extractive multi-document text summarization. Experiments have been performed with Document Understanding Conferences (DUC) datasets, and the model performance has been assessed with eight Recall-Oriented Understudy for Gisting Evaluation (ROUGE) metrics. Even more, the combination formed by different semantic similarity measures obtained the highest results in comparison with the other models.



The main purpose of text summarization systems is to reduce the amount of textual information in a document collection by focusing on the most important topics and avoiding unnecessary portions. One of the summarization techniques named extractive summarization selects the most relevant sentences from the text to provide a representative summary. There are few studies that look at the performance of combining various Semantic similarity and word weighting techniques in automatic text-editing systems.

Fig. 3. An Example of Extractive Summarization of the Proposed Model.

Because each subject has 45 subtopics, we may construct a summary for each of them in our suggested approach. For each of the comparative models and our recommended strategy, we've compiled summaries. Summaries of various sizes were created using different compression ratios of 5 percent, 15%, 25%, and 50% of the original material. The ROUGE-N score metric assesses the quality of the summaries that were created for the purposes of this section.

When it comes to automated summaries, ROUGE is often regarded the gold standard. Rouge contrasts the summaries produced by machines with the summaries created manually (reference summaries). ROUGE-1, ROUGE-2, and ROUGE-L summaries are evaluated at various levels of granularity, giving findings in terms of Precision (P), Recall (R), and F-score (F). For evaluation, ROUGE-1 is used and the results articulated in the following tables are for ROUGE-1. The results of the proposed model on DUC 2007 dataset are articulated in the Tables IV, V and VI.

TABLE IV. AVERAGE F-SCORE ROUGE SCORE VALUES AT COMPRESSION 5% , 15%, 25% RATES

Measure	ROUGE Score (5%)	ROUGE Score (15%)	ROUGE Score (25%)
Tf-idf	0.22	0.23	0.22
Tf-isf	0.22	0.23	0.22
Rtf-sisf	0.23	0.22	0.23
Okapi BM25	0.22	0.24	0.22
Resnik-tf-idf	0.24	0.23	0.24
Resnik-tf-isf	0.24	0.25	0.24
Resnik-Rtf-sisf	0.24	0.25	0.26
Hybrid-Tf-idf	<b>0.26</b>	<b>0.26</b>	<b>0.27</b>

TABLE V. AVERAGE PRECISION ROUGE SCORE VALUES AT COMPRESSION 5%, 15%, 25% RATES.

Measure	ROUGE Score (5%)	ROUGE Score (15%)	ROUGE Score (25%)
Tf-idf	0.23	0.23	0.22
Tf-isf	0.23	0.23	0.22
Rtf-sisf	0.24	0.22	0.23
Okapi BM25	0.23	0.24	0.22
Resnik-tf-idf	0.25	0.23	0.24
Resnik-tf-isf	0.24	0.25	0.24
Resnik-Rtf-sisf	0.26	0.25	0.26
Hybrid-Tf-idf	<b>0.28</b>	<b>0.28</b>	<b>0.28</b>

TABLE VI. AVERAGE RECALL ROUGE SCORE VALUES AT COMPRESSION 5% , 15%, 25% RATES

Measure	ROUGE Score (5%)	ROUGE Score (15%)	ROUGE Score (25%)
Tf-idf	0.24	0.25	0.24
Tf-isf	0.24	0.25	0.24
Rtf-sisf	0.25	0.24	0.25
Okapi BM25	0.24	0.26	0.24
Resnik-tf-idf	0.26	0.25	0.26
Resnik-tf-isf	0.26	0.28	0.26
Resnik-Rtf-sisf	0.26	0.28	0.29
Hybrid-Tf-idf	<b>0.29</b>	<b>0.29</b>	<b>0.30</b>

When compared to the literature, our experimental data has shown that our suggested strategy outperforms the state-of-the-art methodologies, which we feel is important. According to the findings, Table IV further demonstrate that the average Recall values across a variety of variables improve as a consequence of increasing the length of the summary, as can be seen in the tables. Because our model's recall is lower in certain places than it is in others, it is possible that this is due to either a shorter summary or the removal of statistically important characteristics from the model's development process throughout its development. Following an increase in the compression rate from 5 percent to 25 percent, the macro-average F-score values decline somewhat as a consequence of a reduction in the overall accuracy score of the different metrics when the compression rate is raised, according to the study's findings.

However, when comparing the Macro-Averaged F-score values at 22 percent and 23 percent compression rates to the comparative models, as shown in Tables IV, V and VI, the difference is not statistically significant; the difference between the two models is not statistically significant. This demonstrates that the approach given is competitively efficient when compared to the current state of the art. In Table VI, it is shown that, when constructing an average length summary at a 25 percent compression rate, the suggested technique may result in a summary that is more informative than comparison models in certain cases.

This section presented the results of the proposed model and proposed hybrid semantic similarity on word pair similarity and DUC 2007 datasets. The presented results show the efficiency of the model.

## V. CONCLUSION

In light of the vast amount of textual information that is now available, automated extraction text summarization is one of the most widely used methods of organising the data available. Summary techniques make it possible to extract the most important information from a large number of texts in a short amount of time and with minimal effort. When summarizing a text, an extractive summarization method is used that selects the most relevant phrases from the text and presents them in a way that is accurate representation of the text in its entirety. Information extraction systems that extract



information from a large number of documents, such as text summarizing systems, have as their primary objective the reduction of textual information in a document collection. Achieving this is accomplished by concentrating on the most important themes and eliminating any unnecessary information. When it came to developing an automated summary system, the previous study discovered that a variety of strategies, including term-weighting schemes and similarity metrics, were used in the process of development. Currently, there is only a small body of research that examines how different Semantic Similarity and word weighting algorithms perform when used in conjunction with one another in the field of automated text summarization. This study looked at a number of different semantic similarity metrics in the context of extractive multi-document text summary, and we discovered that they were all fairly accurate in terms of similarity. This research looked into different semantic similarity metrics that could be used in extractive multi-document text summarization, and the results were published. Various ROUGE criteria were used to evaluate the model's performance in this study, which was carried out using DUC dataset. When the results of the various semantic similarity metrics were combined, the resulting model produced the most favourable results when compared to the other models in this study.

#### REFERENCES

- [1] A. K. Srivastava, D. Pandey, and A. Agarwal, "Extractive multi-document text summarization using dolphin swarm optimization approach," *Multimedia Tools and Applications*, vol. 80(7), pp. 11273-11290, 2021.
- [2] W. S. El-Kassas, C. Salama, A. A. Rafea, and H. K. Mohamed, "Automatic text summarization: A comprehensive survey," *Expert Systems with Applications*, vol. 165, pp. 113679, 2021.
- [3] L. Dong, M. N. Satpute, W. Wu, and D. Z. Du, "Two-phase multidocument summarization through content-attention-based subtopic detection," *IEEE Transactions on Computational Social Systems*, vol. 8(6), pp. 1379-1392, 2021.
- [4] J.M. Sanchez-Gomez, M. A. Vega-Rodríguez, and C. J. Perez, "Parallelizing a multi-objective optimization approach for extractive multi-document text summarization," *Journal of Parallel and Distributed Computing*, vol. 134, pp. 166-179, 2019.
- [5] J.M. Sanchez-Gomez, M. A. Vega-Rodríguez, and C. J. Perez, "A decomposition-based multi-objective optimization approach for extractive multi-document text summarization," *Applied Soft Computing*, vol. 91, pp. 106231, 2020.
- [6] M. Bidoki, M. R. Moosavi, and M. Fakhrahmad, "A semantic approach to extractive multi-document summarization: Applying sentence expansion for tuning of conceptual densities," *Information Processing & Management*, vol. 57(6), pp. 102341, 2020.
- [7] B. Mutlu, E. A. Sezer, and M. A. Akcayol, "Candidate sentence selection for extractive text summarization," *Information Processing & Management*, vol. 57(6), pp. 102359, 2020.
- [8] H. K. Thakkar, P. K. Sahoo, and P. Mohanty, "DOFM: Domain Feature Miner for robust extractive summarization," *Information Processing & Management*, vol. 58(3), pp. 102474, 2021.
- [9] A. Hernández-Castañeda, R. A. García-Hernández, Y. Ledeneva, and C. E. Millán-Hernández, "Language-independent extractive automatic text summarization based on automatic keyword extraction," *Computer Speech & Language*, vol. 71, pp. 101267, 2022.
- [10] K. Yao, L. Zhang, T. Luo, and Y. Wu, "Deep reinforcement learning for extractive document summarization," *Neurocomputing*, vol. 284, pp. 52-62, 2018.
- [11] A. Joshi, E. Fidalgo, E. Alegre, and L. Fernández-Robles, "SummCoder: An unsupervised framework for extractive text summarization based on deep auto-encoders," *Expert Systems with Applications*, vol. 129, pp. 200-215.
- [12] L. Cagliero, and M. La Quatra, "Extracting highlights of scientific articles: A supervised summarization approach," *Expert Systems with Applications*, vol. 160, pp. 113659, 2020.
- [13] S. Lamsiyah, A. El Mahdaouy, B. Espinasse, and S. E. A. Ouatik, "An unsupervised method for extractive multi-document summarization based on centroid approach and sentence embeddings," *Expert Systems with Applications*, vol. 167, pp.114152, 2021.
- [14] J.M. Sanchez-Gomez, M. A. Vega-Rodríguez, and C. J. Perez, "The impact of term-weighting schemes and similarity measures on extractive multi-document text summarization," *Expert Systems with Applications*, vol. 169, pp. 114510, 2021.
- [15] M. Mojriari, and S. A. Mirroshandel, "A novel extractive multi-document text summarization system using quantum-inspired genetic algorithm: MTSQIGA," *Expert systems with applications*, vol. 171, pp. 114555, 2021.
- [16] A. Khurana, and V. Bhatnagar, "Investigating Entropy for Extractive Document Summarization," *Expert Systems with Applications*, vol. 187, pp. 115820, 2022.
- [17] R. Rani, and D. K. Lobiyal, "A weighted word embedding based approach for extractive text summarization," *Expert Systems with Applications*, vol. 186, pp. 115867, 2021.
- [18] M. Tomer, and M. Kumar, "Multi-document extractive text summarization based on firefly algorithm," *Journal of King Saud University-Computer and Information Sciences*, vol. 34(8), pp.6057-6065, 2021.
- [19] P. Verma, A. Verma, and S. Pal, "An approach for extractive text summarization using fuzzy evolutionary and clustering algorithms," *Applied Soft Computing*, vol. 120, p.108670, 2022.
- [20] J. Chen, and H. Zhuge, "Extractive summarization of documents with images based on multi-modal RNN," *Future Generation Computer Systems*, vol. 99, pp. 186-196, 2019.
- [21] Resnik, "Using information content to evaluate semantic similarity in a taxonomy," *arXiv preprint cmp-lg/9511007*, 1995.
- [22] Harispe, Ranwez, Janaqi, and Montmain, "Semantic similarity from natural language and ontology analysis," *Synthesis Lectures on Human Language Technologies*, vol. 8(1), pp. 1-254, 2015.
- [23] Jiang, and David W Conrath, "Semantic similarity based on corpus statistics and lexical taxonomy," *arXiv preprint cmp-lg/9709008*, 1997.
- [24] Lin, and Dekang, "An information-theoretic definition of similarity," *Proceedings of ICML*, vol. 98, 1998.
- [25] Leacock, Claudia, and Martin Chodorow, "Combining local context and WordNet similarity for word sense identification," *WordNet: An electronic lexical database*, vol. 49(2), pp. 265-283, 1998.
- [26] Wu, Zhibiao, and Martha Palmer, "Verbs semantics and lexical selection," *Proc. Asso. Comp. Ling*, 1994.
- [27] Rada, Roy, et al, "Development and application of a metric on semantic nets," *IEEE trans. on syst., man, and cyber*. vol. 19.1, pp.17-30, 1989.
- [28] Li, Bandar and McLean, "An approach for measuring semantic similarity between words using multiple information sources," *IEEE Trans.on know. and data eng.*, vol. 15(4), pp. 871-882, 2003.
- [29] Rubenstein, Herbert, and John B. Goodenough, "Contextual correlates of synonymy," *Commun. ACM*, vol. 8(10), pp. 627-633, 1965.
- [30] Manh, Hai Cao, Huong Le Thanh, and Tuan Luu Minh, "Extractive Multi-document Summarization using K-means, Centroid-based Method, MMR, and Sentence Position," *Proceedings of the Tenth International Symposium on Information and Communication Technology*, pp. 29-35, 2019.
- [31] M. Krishna Siva Prasad, and Poonam Sharma, "Similarity of Sentences With Contradiction Using Semantic Similarity Measures," *The Computer Journal* 65, no. 3 (2022): 701-717.
- [32] M. Krishna Siva Prasad, and Poonam Sharma, "Exploring intrinsic information content models for addressing the issues of traditional semantic measures to evaluate verb similarity." *Computer Speech & Language* 71 (2022): 101280.

# An Efficient Hybrid LSTM-CNN and CNN-LSTM with GloVe for Text Multi-class Sentiment Classification in Gender Violence

Abdul Azim Ismail<sup>1</sup>

Faculty of Computer and Mathematical Science  
Universiti Teknologi MARA  
Shah Alam, Malaysia

Marina Yusoff<sup>2</sup>

Institute for Big Data Analytics and Artificial Intelligence  
(IBDAAD), Kompleks AI-Khawarizmi  
Universiti Teknologi MARA  
Shah Alam, Malaysia

**Abstract**—Gender-based violence is a public health issue that needs high concern to eliminate discrimination and violence against women and girls. Several cases are through the offline organization and the respective online platform. However, many victims share their experiences and stories on social media platforms. Twitter is one of the methods for locating and identifying gender-based violence based on its type. This paper proposed a hybrid Long Short-Term Memory (LSTM) and Convolution Neural Network CNN with GloVe to perform multi-classification of gender violence. Intimate partner violence, harassment, rape, femicide, sex trafficking, forced marriage, forced abortion, and online violence against women are eight gender violence keyword for data extraction from Twitter text data. Next is data cleaning to remove unnecessary information. Normalization converts data into a structure the machine can recognize as model input. The evaluation considers cross-entropy loss parameters, learning rate, an optimizer, and epochs. LSTM+GloVe vector embedding outperforms all other methods. CNN-LSTM+Glove and LSTM-CNN+GloVe achieved 0.98 for test accuracy, 0.95 for precision, 0.94 for recall, and 0.95 for the f1-score. The findings can help the public and relevant agencies differentiate and categorize different types of gender violence through text. With this effort, the government can use as one of the mechanisms that indirectly can support monitoring of the current situation of gender violence.

**Keywords**—Gender-based violence; deep learning; convolution neural network; long short-term memory; convolution neural network - long short-term memory; long short-term memory - convolution neural network; global vector; multi-class text classification

## I. INTRODUCTION

GBV is a worldwide public health concern [1]. GBV refers to any violence toward any individual because of the individual's gender [2]. One-third of women have experienced sexual or physical violence [3]. GBV is a type of violence perpetrated against women and girls. It can physically, sexually, and mentally injure women and girls through violence, compulsion, or arbitrary denial of liberty. The Sustainable Development Goals sought to eliminate gender discrimination and violence against women and girls [4]. As a result, everyone should feel safe at home or in public, especially women who may be victims of violence.

For example, an actress, resorted to social media to expose her experiences with sexual harassment in Hollywood. The public's focus on this issue has increased awareness of GBV, particularly sexual harassment [5]. Meanwhile, a Malaysian woman resorted to Twitter to complain about harassment using an e-hailing service [6]. These stories raise public consciousness. However, online social media allows disaffected people to control specific people's lives and utilize the anonymity or social distancing afforded by the internet to harass others [7]. Sexting the other sex, for example, is one of the most divisive issues on social networking. The evidence leads to sexual harassment and mental health problems [8].

People who seek to harass women and advocate violence against women can do so anonymously through social media platforms [9]. This campaign primarily targets female public figures, including politicians, journalists, and public figures [10]. Consequently, measures must be taken to address the seemingly endless instances of gender-based violence. Additionally, domestic violence instances are underreported, with the police, the health care system, and non-governmental organizations saying that just 7 percent of victims sought assistance from these institutions [11]. The principal perpetrators face stigma and societal pressures [12]. The fifth Sustainable Development Goal (SDG) seeks to eliminate all types of prejudice and violence. As a direct consequence of this, these challenges require attention.

Social media to collect data for a study on gender violence. On the other hand, a study utilizing 0.7 million tweets and a deep learning system discovered that sexual assaults are more likely to be performed by someone who knows than by someone who does not know [13]. Researchers also used Twitter data to construct a detection tool for sexual harassment and cyberbullying using machine learning and frequency inversion document frequency (TF-IDF) [14]. In addition, one study analyzed patient anecdotes about their healthcare experiences using topic modeling with Latent Dirichlet Allocation (LDA) and sentiment analysis on Twitter data [15]. As a result, this research aims to conduct a text classification that can separate the meaning of GBV-related text content. This study improves the current method for managing violent content on social media, namely the detection of Gender-Based Violence.

The following are the significant contributions of the subsequent paper:

- This study data collection is from Twitter, the public data on social media related to gender violence issues during the Covid-19 pandemic from January 01, 2022, until April 01, 2022, worldwide.
- The proposed model of deep learning classifier Convolutional Neural Network-Long Short-Term Memory with GloVe (CNN-LSTM+GloVe) and LSTM-CNN+GloVe applies to sentiment analysis for gender violence.
- The comparative analysis of different deep learning and hybrid classifiers with the suggested model verifies its performance.

The section is organized in the following manner throughout the rest of the paper: Section II goes over the connected works. The materials and methods are in Section III. Section IV presents the results of the experiment. The discussion offered in Section V and Section VI constitutes the study's conclusion.

## II. RELATED WORKS

A text categorization method is a supervised machine learning in which unstructured text assign to specified categories. Text categorization aids in the organization, structuring, and classification of text documents such as Twitter data, news articles, and medical records. The process of text classification is appropriate for extracting new information from a textual source [16]. The study looks at how text classification identifies gender violence as one of the text's features. They seek occurrences of violence in social media data using text categorization in Arabic dialect. One of the objectives of this study will be to evaluate different text classification methods. This study uses supervised machine learning techniques such as support vector machine (SVM), K-nearest neighbors (KNN), and Bayesian boosting with complement naive Bayes to extract information from 700,000 tweets. The hashtag #Metoo appears in these messages. According to him, there is a scarcity of studies that use Arabic for data analysis. As a result, additional research is required.

Using text classification algorithms, investigating domestic violence in intimate relationships to grasp the clinical importance of the victim is better accomplished by using a technique known as a "word cloud," which sorts text based on Python scripts [17]. This study's primary source of information was the Rio Grande do Sul Legal Medical Department. Based on the findings of this study, they concluded that using a word cloud to assess a variety of topics presented by participants was feasible. Despite this, they emphasized the need for more

significant research into the applicability of these techniques [18]. According to their research findings, this work recognized GBV messages on social media using BERT and NLP. This study evaluates the material to determine whether it was aggressive or peaceful.

The researchers discovered that after incorporating a preprocessing step in the initial dataset, the area under the curve, accuracy, sensitivity, and specificity for a total of 16421 messages were, respectively, 0.9603, 0.8909, 0.8826, and 0.8989. Overall, their findings indicated that the categorization performance of their text dataset was satisfactory [18]. A study that used the Latent Dirichlet Allocation method on Twitter data produced roughly 56% coherence and 18 ambiguities [19]. The coherence and complexity scores look to be excellent, but there is an opportunity for development to attain even higher outcomes. Based on the findings, it can be inferred that many studies on gender violence and social media have been conducted. One connected study uses gender violence data from Twitter to classify the corpus using text classification based on previously labeled data.

Furthermore, Khatua et al. used Twitter data to build a multilayer perceptron (MLP), convolutional neural network (CNN), long short-term memory (LSTM), and bidirectional LSTM, all of which are similar to the approaches outlined in this study (Bi-LSTM) [13]. Their study examined the many types of sexual violence and the associated hazards. Between October 15, 2017, and October 26, 2017, they collected 0.7 million tweets using the hashtag #Metoo. CNN, LSTM, and bi-LSTM achieve precisions of 0.83, 0.82, and 0.81 during the text classification process, whereas MLP achieves a precision of 0.77. CNN has the highest accuracy of the four algorithms; moreover, all have an accuracy of less than 0.90, improving with ongoing research. CNN has the highest level of accuracy. According to the text categorization research, it is conceivable to undertake an additional study on deep learning algorithms such as CNN, LSTM, and the hybrid LSTM-CNN technique.

## III. MATERIALS AND METHODS

This section describes the study's structure, method, and procedure. A few steps of this work adapted Offer's approach [20]. This study methodology includes data collection, preprocessing, feature extraction, and modeling. Twitter text data is scraped using Twitter Intelligence Tool (Twint). After scraping, the dataset is preprocessed to remove text noise. The training set will be labeled by GBV dataset. The dataset is then used to generate training and testing sets. The model's training process uses the CNN, LSTM, and LSTM-CNN machine learning algorithms. If a text does or does not contain GBV can be predicted using the testing set, which is not labeled. The conceptual framework for the research is shown in Fig. 1.

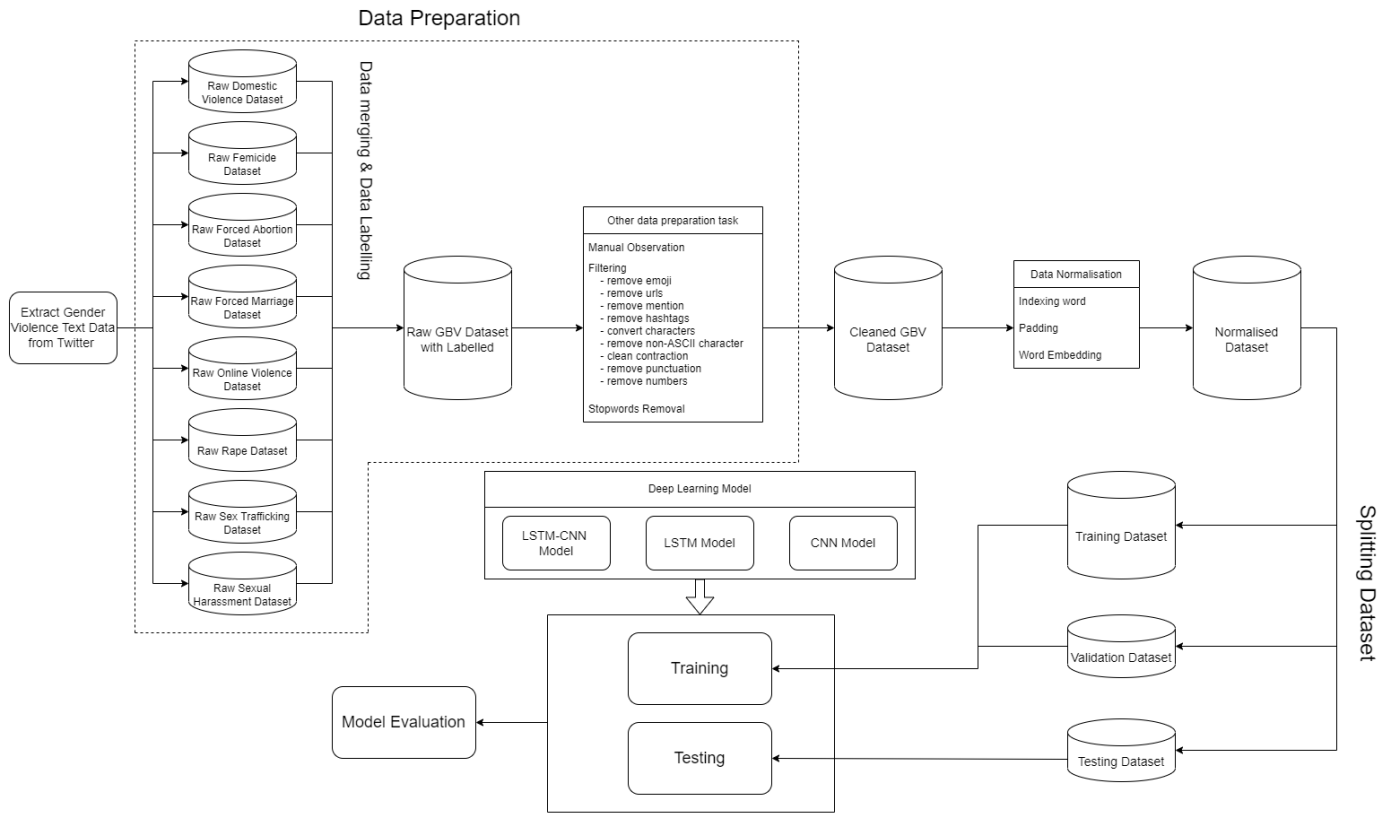


Fig. 1. Research Framework.

A. Data Preparation

This section briefly explains the steps in data preparation, including data acquisition, data labeling, data merging, manual observation, filtering, and stopwords.

1) *Data acquisition*: In this phase, the research data scrape from the web. We use Twitter Intelligence Tools (Twint) to extract tweets based on a keyword. It is a Python web scraping program that allows users to scrape tweets without limitations, considering that it does not use Twitter. This research requires many documents or results relating to Twitter's unlimited API, which only delivers 3200 tweets each. An open-source tool with various features. The total data gathered for each category of gender violence are 300000. The keyword used to extract the data is in Table I. We determined eight categories of GBV, which are domestic violence, sexual harassment, rape, femicide, sex trafficking, forced marriage, forced abortion, and female genital mutilation [21][22].

2) *Data labeling*: Labeling annotates every tweet in the dataset with appropriate classes. All tweets in the dataset into eight GBV classifications to create multi-class data.

3) *Data merging*: Data merging involves combining the obtained datasets into a single dataset from eight datasets representing eight categories of GBV.

4) *Manual observation*: Recall (R) is a combination of all objects grouped into a specific class. The formula of recall is in Eq. 4.

TABLE I. TYPE OF GBV BASED ON KEYWORDS

Class	Keywords
Domestic Violence	Intimate partner violence, domestic violence, domestic abuse
Sexual Harassment	Sexual harassment, harassment, stalking
Rape	Rape, rape culture, corrective rape
Femicide	Femicide, femicide, honor killing, honour killing
Sex Trafficking	Sex trafficking
Forced Marriage	Forced marriage, child marriage
Forced Abortion	Forced abortion, forced sterilization, coerced sterilization, unwanted sterilization, forced miscarriage
Female Genital Mutilation	female genital mutilation, female circumcision, female genital cutting

Manual observation can refer to an individual's observation of certain things or works. Typos or grammatical errors and Unwanted data from the dataset may include text report articles and duplicated content. Meanwhile, features such as location, language, mentions, and URLs are unimportant to the research because they provide no meaningful information or value to the study.

5) *Filtering*: Several undesired things inside the phrases during the manual observation procedure can be deemed noise to the dataset. As a result, the filtering process will remove all of the extraneous noise within the corpus, such as emojis, URLs, mentions, and hashtags. It is necessary to lower the dataset dimensions and improve the learning process.

6) *Stopwords*: Stopwords are commonly used in a text mining project with little influence [23]. "The", "A", "Is," and "Are" are stop words. Stop words removed to lower the document's high dimensionality and computing time. Before filtering, each data set had 107 words; after, it had 52. Fewer words will lead to a faster calculation.

### B. Data Normalization

Before the dataset can be applied to the deep learning model, it must undergo a data normalization procedure. It is to verify that the dataset is in the same format or condition, particularly for text data, which will be the primary component of the training process. The four primary processes are the word indexing procedure, padding, word embedding, and one-hot encoding in this study's dataset.

1) *Word indexing*: As the machine does not understand words, it converts them to integers. This study uses Keras library functions fit on texts and sequence on texts.

2) *Padding*: This study used padding to standardize each dataset's text data length. Due to the dataset's varied text lengths, this procedure is essential so that it may be model input. First, we require the dataset's maximum length. Set all text properties to the same length.

3) *Word embedding*: This study uses GloVe embedding to perform a pre-trained word-vector model. The GloVe has educated 2 billion tweets and 1.2 million vocabularies. (R) is a combination of all objects grouped into a specific class. The formula of recall is in Equation 4.

4) *One-Hot encoding*: Each tweet's class attribute is hot encoded. It converts categorical data into 1 and 0 classes. 1 represents this category, 0 otherwise.

### C. Splitting Datasets

The training dataset comprises 80% of the total, whereas the testing dataset will comprise 20%. This project employs supervised learning. As a result, we require validation.

### D. Proposed Model

In this phase, constructing and implementing a deep learning model will be done. The deep learning model that will be used is the convolutional neural network (CNN), long-short term memory (LSTM), LSTM-CNN, and CNN-LSTM. Thus, in this section, the model's architecture will be discussed. Fig. 2 illustrates the model architecture for all four models.

1) *CNN*: CNN's deep learning model is popular. Fig. 2 shows that the model will accept input at the embedding layer.

The convolution layer extracts features and generates feature maps. The pooling layer shrinks feature maps. The first dense layer utilized the "relu" activation function, second layer used "softmax" Output is text type or topic prediction. In this design, the embedding layer translated input into embedding vectors before delivering them to LSTM. Each LSTM cell in the LSTM layer took each embedding vector, determined the relevant information, and formed a new encoding vector. Two dense layers would assist in increasing the class categorization based on input vector attributes. The first dense layer utilized the "relu" activation function. The second layer used the "softmax".

2) *LSTM*: In this design, the embedding layer transformed the input into a sequence of embedding vectors before sending them to the LSTM layer. Each LSTM cell in the LSTM layer took each embedding vector, selected the critical information that needed to be maintained, and then generated a new encoding vector based on the previously stored information. Two dense layers to improve class categorization based on the features gathered from the input vectors. The first dense layer utilized the activation function "relu," while the second layer used the activation function "softmax" to predict the output.

3) *Hybrid CNN-LSTM*: In this setup, initially, the embedding layer converted the input phrases into embedding vectors. Once the embedding vector is received, the convolution layer produces feature maps by extracting features. The pooling layer will help to reduce the feature maps. Next, the LSTM layer took the output of the convolutional layer and selected the critical information to be maintained. Then a new encoding vector based on the previous information will be stored. Lastly, two dense layers will help to improve the class categorization. (R) is a combination of all objects grouped into a specific class. The formula of recall is in Equation 4.

4) *Hybrid LSTM-CNN and CNN-LSTM*: In this configuration, the embedding layer first turned the input phrases into embedding vectors before the model could begin to run. After receiving each embedding vector, the LSTM layer learned the words in order, stored them, and created a new encoding vector. The convolution layer processes the output and creates a series of feature maps, which are subsequently combined by the pooling layer. Two dense layers increase class categorization based on input vector attributes. The first dense layer employed "relu" and the second layer considered a dataset to predict the output. As a result, the training dataset is divided by 9:1, with 90% remaining as training and 10% retraining and validation. Fig. 2 illustrates the overall model architecture of CNN, LSTM, CNN-LSTM, and LSTM-CNN models. (R) is a combination of all objects grouped into a specific class. The formula of recall is in Eq. 4.

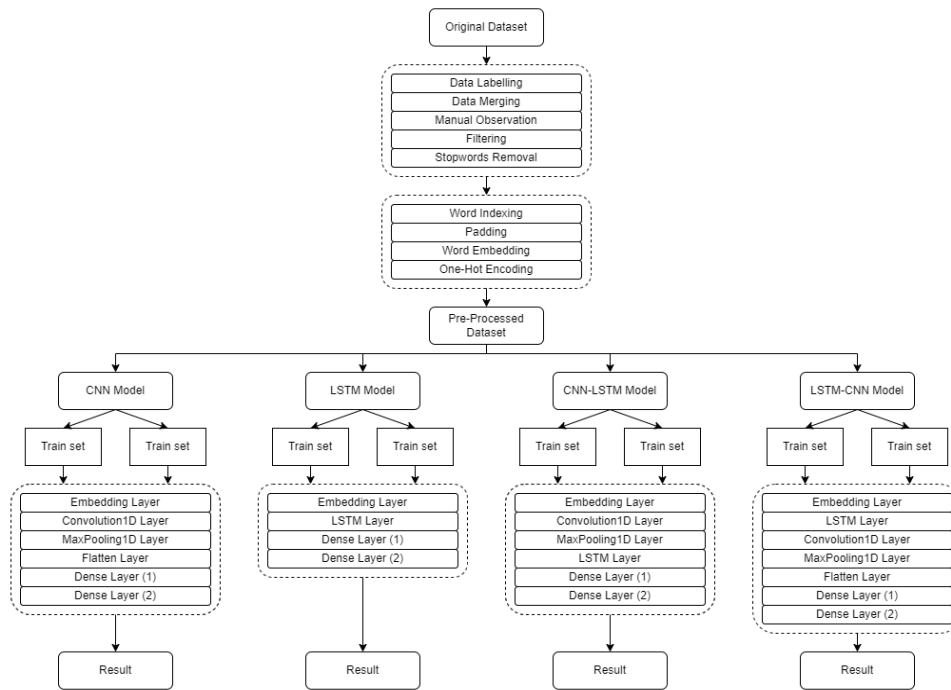


Fig. 2. Model Architecture.

### E. Model Evaluation

Supervised learning involves training and testing to find the optimum model for training accuracy, loss, and computational time confidence. Total predictions divided by accurate predictions is model accuracy. Accuracy increases model performance. Equation (1) calculates accuracy. (R) is a combination of all objects grouped into a specific class. The formula of recall is in Equation (4).

$$accuracy = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}} \quad (1)$$

Loss is the difference between the actual value of the issue and what the model forecasts. The less accurate the model, the more significant the loss. A categorical cross entropy function calculates loss. Thus, Eq. (2) shows loss evaluation.

$$Loss = -\sum_{i=1}^{\text{output size}} y_i \times \log \hat{y}_i \quad (2)$$

where output size is the number of scalar values in the model output,  $y_i$  is the goal value, and  $\hat{y}_i$  is the  $i$ -th scalar value in the model output. A testing technique predicts the trained model's correctness to determine its accuracy. After the modeling phase, CNN, LSTM, and hybrid LSTM-CNN performance will be evaluated. Precision, recall, and f1-score can evaluate text classification performance [24]. Precision (P) estimates the ratio of the true positives among the cluster. The formula of precision is in Eq. 3.

$$Precision = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} \quad (3)$$

Recall (R) is a combination of all objects grouped into a specific class. The formula of recall is in Eq. 4.

$$Recall = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} \quad (4)$$

F-measure (F) is a combination of precision and recall that measures the cluster that contains only objects of a particular class and is used to balance false negatives by weighting recall parameter  $\eta \geq 0$ . The formula of the F-measure is in Eq. 5.

$$F\text{-measure} = \frac{(2 \times \text{Precision} \times \text{Recall})}{(\text{Precision} + \text{Recall})} \quad (5)$$

To calculate these performance indicators, we need a confusion matrix of the model. True positive (TP) describes how well the model predicts the class. True negative (TN) means the model predicts it to be false. False positive means the model inaccurately predicts the true statement or class, while false negative means the opposite. The illustration is different in a multi-class classification with more than two labels. Thus, Fig. 3 depicts a confusion matrix with more than two classes [24].

		Predicted Class			
		C <sub>1</sub>	C <sub>2</sub>	...	C <sub>N</sub>
Actual Class	C <sub>1</sub>	C <sub>1,1</sub>	FP	...	C <sub>1,N</sub>
	C <sub>2</sub>	FN	TP	...	FN
	...	...	...	...	...
	C <sub>N</sub>	C <sub>N,1</sub>	FP	...	C <sub>N,N</sub>

Fig. 3. Multi-class Confusion Matrix.



#### IV. EXPERIMENTAL RESULTS

##### A. Datasets

The experiment used Twitter text as primary data based on the eight gender violence categories. The total tweets extracted are 103 197 English tweets from around the world. The dataset has undergone data preprocessing and cleaning, including manual observation, filtering, stop word removal, and normalization. After preprocessing and cleaning, there were 85,697 tweets. The class attributes in this dataset need to be balanced. This dataset is homogeneous as it only contains string values after preprocessing and cleaning.

##### B. Parameter Settings

This subsection explains CNN, LSTM, CNN-LSTM, and LSTM-CNN model parameters. Four models employ essentially constant parameters. The complete experiment's embedding dimension is 100 since the GloVe pre-trained embedding dimension is 100. The data set shows 61766 words. However, we account for one vacant space. The long sentences in the dataset are 42 words; hence in this experiment, the maxlen parameter is set at 42. CNN has 100 filters. LSTM's hidden layer is 100. Max Pooling is utilized as the pooling layer because it is frequent in deep learning models. This research will implement two dense layers: the first will employ 100-dimensional Relu activation, while the second will use 8-dimensional Softmax activation. Next, we use Adam as the model's optimizer with a learning rate of 0.0003. Set 20 epochs. Table II lists parameters.

TABLE II. PARAMETER SETTING

Parameter	Parameter Value
Embedding Dimension	100
Number of words (unique)	61767
Maxlen	42
Pooling	Max Pooling
Dense (1)	Activation = 'relu', dimension = 100
Dense (2)	Activation = 'softmax', dimension = 8
loss	categorical_crossentropy
Learning rate	0.0003 @ 3e-4
optimizer	Adam
Validation split	0.1
Epoch number	20
Word embedding	With GloVe and without embedding

##### C. Experimental Results

The study features two experiments using GloVe and without GloVe. The analysis will be based on experiments on

the four models, comparing their performance in training and testing.

1) *Training result without GloVe*: The accuracy and loss learning curves of the CNN, LSTM, LSTM-CNN, and LSTM-CNN are depicted in Fig. 4(a), (b), (2), (d), (e), (f), (g), and (h). LSTM, as shown in Fig. 4(c), has the smallest gap in accuracy and measurement compared to other models. The same result for loss value. The hybrid LSTM-CNN and CNN-LSTM, on the other hand, outperform a single CNN in terms of accuracy. LSTM has training and validation accuracy of around 0.3418 to 0.9995 and 0.6412 to 0.9781, respectively, while training and validation loss is 0.3031-0.049133. and 0.1823 to 0.0287.

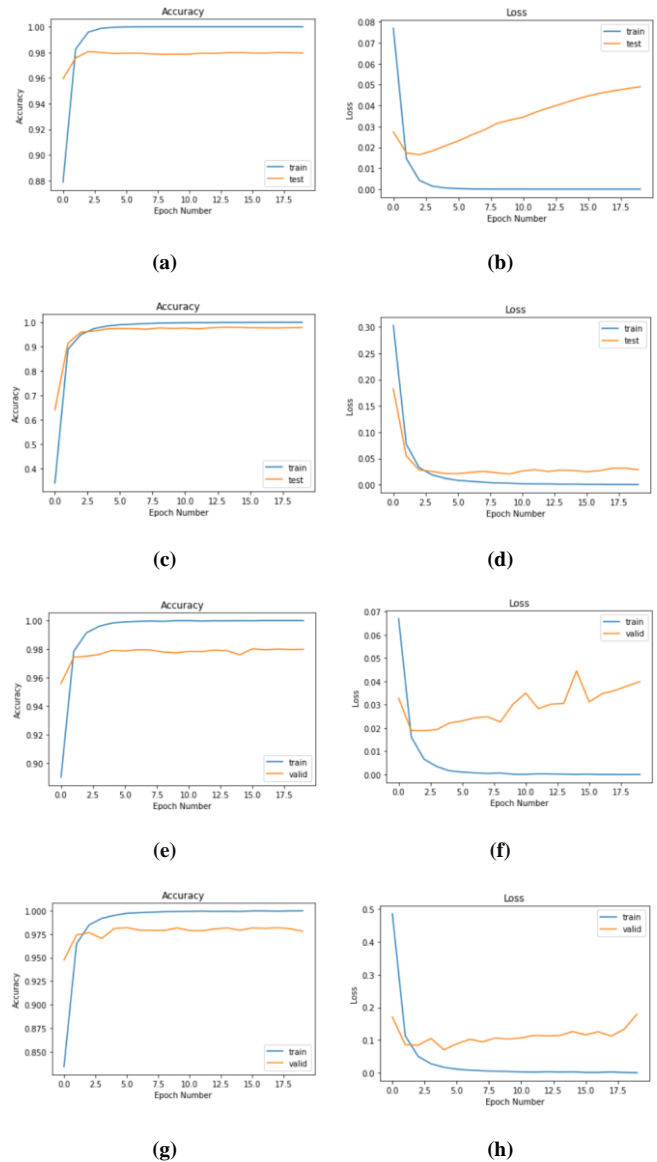


Fig. 4. Learning Curves Graphs (a) Accuracy for CNN (b) Loss for CNN (c) Accuracy for LSTM (d) Loss for LSTM (e) Accuracy for LSTM-CNN (f) Loss for LSTM-CNN (g) Accuracy for CNN-LSTM (h) Loss for CNN-LSTM.

2) *Testing results without GloVe*: Table III displays the testing performance of the models using the confusion matrix accuracy and loss measure of the multi-classes Twitter text data. Overall, all models receive a comparable categorization score. This demonstrates that for most classes, the TP value is more notable than the FP and FN scores, except for the femicide class, whose TP and FP+FN scores are comparable. Overall, CNN and LSTM memory scores for domestic violence and rape are 0.99, whereas femicide scores are 0.80. The f1 score for domestic violence, rape, and sex trafficking is 0.99. TP is superior to FP and FN for both CNN-LSTM and LSTM-CNN. Overall, the model achieves an accuracy of 0.981 with a loss of 0.039 on the testing dataset. According to the model, sex trafficking, sexual harassment, and femicide had an

accuracy of 0.99 and 0.85, respectively. Domestic violence and rape have a recall rate of 0.99, while femicide is 0.77. The f1 score for domestic violence, rape, and sexual harassment was 0.99.

On the testing dataset, CNN's femicide class achieves a precision of 0.982% with a loss of 0.048. According to the model, forced abortion has a precision of 1.00, while femicide has a precision of 0.78. However, the femicide class has the fewest records in the test dataset of 235. This class yields comparable results for LSTM, CNN-LSTM, and LSTM-CNN models. Based on the test set, we can predict that the label with the most significant number of datasets will have the highest scores for performance metrics.

TABLE III. RESULT OF CNN, LSTM, CNN-LSTM, AND LSTM-CNN WITHOUT GLOVE

Label	CNN+GloVe						LSTM+GloVe						LSTM-CNN+GloVe						LSTM-CNN+GloVe					
	P	R	F1	S	Loss	Acc	P	R	F	S	Loss	Acc	P	R	F1	S	Loss	Acc	P	R	F1	S	Loss	Acc
Domestic Violence	0.99	<b>0.99</b>	<b>0.99</b>	3577	0.05	0.98	<b>0.99</b>	0.98	<b>0.99</b>	3577	0.03	0.98	0.98	<b>0.99</b>	<b>0.99</b>	3577	0.04	0.98	0.99	0.98	0.98	3577	0.15	0.97
Femicide	0.78	0.80	0.79	235			0.73	0.80	0.76	235			0.85	0.77	0.81	235			0.76	0.76	0.76	235		
Forced Abortion	<b>1.00</b>	0.94	0.97	285			0.96	0.96	0.96	285			0.94	0.97	0.95	285			0.97	0.97	0.97	285		
Forced Marriage	0.98	0.98	0.98	584			0.97	0.97	0.97	584			0.98	0.98	0.98	584			0.97	0.98	0.98	584		
Online Violence	0.96	0.86	0.90	253			0.89	0.85	0.87	253			0.93	0.89	0.91	253			0.84	0.90	0.87	253		
Rape	0.98	<b>0.99</b>	<b>0.99</b>	5439			0.98	<b>0.99</b>	<b>0.99</b>	5439			0.98	<b>0.99</b>	<b>0.99</b>	5439			0.97	0.99	0.98	5439		
Sex Trafficking	0.99	0.98	<b>0.99</b>	1369			<b>0.99</b>	0.98	<b>0.99</b>	1369			<b>0.99</b>	0.98	0.98	1369			0.98	0.97	0.97	1369		
Sexual Harassment	0.98	0.98	0.98	5041			0.98	0.98	0.98	5041			<b>0.99</b>	0.98	<b>0.99</b>	5041			0.99	0.97	0.98	5041		
Average	0.96	0.94	0.95	1678	0.94	0.94	0.94	1678	0.96	0.94	0.95	1678	0.93	0.94	0.94	1678								

3) *Training result with GloVe*: The training results for CNN+GloVe, LSTM+GloVe, LSTM-CNN+GloVe, and CNN-LSTM+GloVe models are in Fig. 5(a)-5(h). Fig. 5(a) depicts CNN's training accuracy over 20 epochs. The models' training accuracy ranges from 0.8923 to 1.0000, and validation from 0.9595 to 0.9756. It suggests a positive pattern in which both the training and validation sets produced strong results, but there is a significant generalization gap between the two sets. The training loss ranges from 0.0798 to 0.000046438. Fig. 5(b) shows that the validation loss began at 0.0333 and stopped at 0.0394.

Fig. 5(c) shows that the LSTM+GloVe training accuracy is 0.6558 to 0.9906, and its validation accuracy is 0.9146 to 0.9830. It suggests a positive pattern in which the training and validation sets produced good results with a small generalization gap. It is worth noting that the training loss begins at 0.1888 and finishes at 0.0064. The validation loss started at 0.0561 and terminated at 0.0127. According to the data and graph in Fig. 5(d), training and validation loss exhibit a decreasing pattern with a minimal generalization gap at the end of training. In terms of training and validation accuracy and loss pattern, the LSTM-CNN+GloVe and CNN-LSTM+GloVe appear to follow a similar trend. LSTM-CNN+GloVe, on the other hand, offers training accuracy that starts at 0.9141 and ends at 0.9994, while validation accuracy starts at 0.9607 and ends at 0.9815. It suggests a positive pattern in which both the training and validation sets produced good results, and there is a large generalization gap between the two sets. The training loss ranges from 0.0582 to 0.00069372. The validation loss started at 0.0267 and finished at 0.0265. Training loss shows a decreasing pattern based on the data and graph. However, validation loss shows an increasing tendency. CNN-LSTM+GloVe produced comparable results.

Table IV shows that the CNN+GloVe's accuracy is 0.976 with a loss of 0.040. Domestic abuse and sex trafficking have the highest precision (0.99), whereas femicide has the lowest (0.62). On recall, domestic violence has 0.99, and femicide is 0.63. The f1-score gives domestic violence 0.99. LSTM shows that the model achieves a 0.983 accuracy value with a 0.013 loss on the testing dataset. The model's average precision value is 0.95, with most labels achieving 0.99 and femicide showing 0.72. For recall, almost all labels score above 0.94, where 0.99 is the highest and 0.68 is the lowest, where seven out of eight label f1 scores average 0.95.

LSTM-CNN+GloVe and CNN-LSTM+GloVe have acceptable results since the TP value is more than FP and FN. FP and FN are higher than TP for just femicide. Table IV indicates that the model achieves a 0.981 accuracy value with a 0.039 loss. The model finds that the average precision value is 0.94, with forced abortion achieving the highest precision (1.00) and femicide the lowest (0.60). Most labels indicate a positive recall above the average of 0.95, where the highest score has been 0.99 and the lowest is 0.73. Seven of eight label ratings are above average for the f1-score (0.94). Meanwhile,

the lowest performance is the femicide since it is the one that has the least number of test datasets with only 235 records.

4) *Results based on computational time*: Table IV demonstrates the computational time that was recorded from the highest training accuracy and loss value. CNN+GloVe, LSTM+GloVe, LSTM-CNN+GloVe, and CNN-LSTM+GloVe recorded more than one hour compared to the models without GloVe. The minimum computational time consumed by CNN+GloVe of about 9 minutes and 10 s; meanwhile, the maximum is LSTM-CNN of about one h 59 min 27 s.

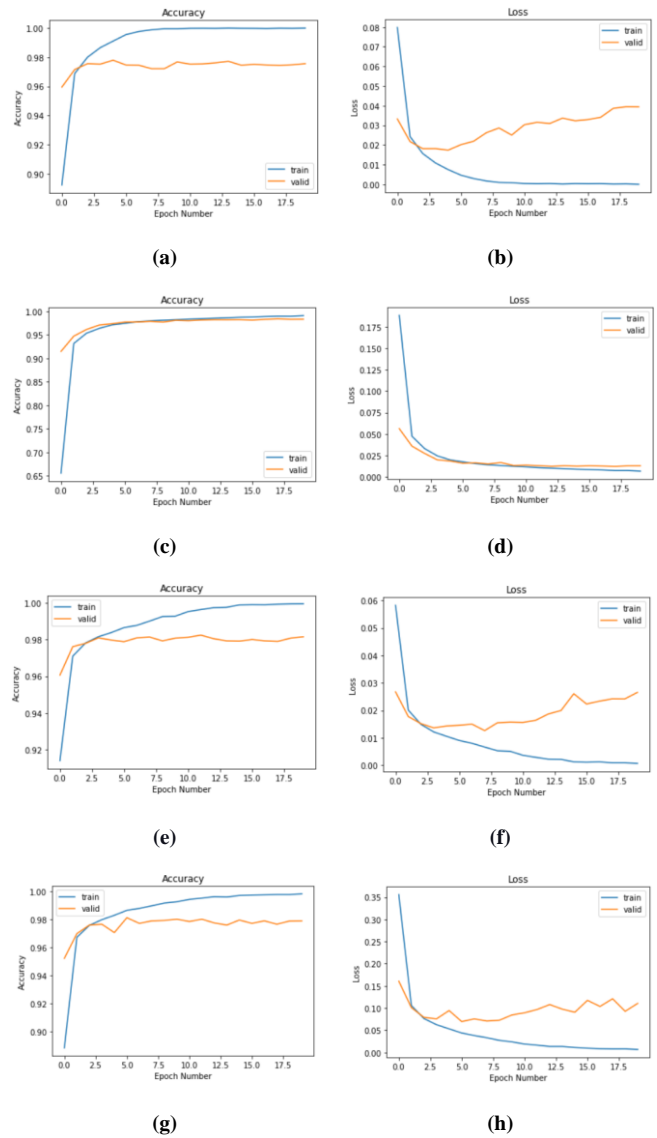


Fig. 5. Learning Curves Graphs (a) Accuracy for CNN+GloVe (b) Loss for CNN+GloVe +GloVe (c) Accuracy for LSTM (d) Loss for LSTM+GloVe (e) Accuracy for LSTM-CNN+GloVe (f) Loss for LSTM-CNN +GloVe (g) Accuracy for CNN-LSTM+GloVe (h) Loss for CNN-LSTM+GloVe.

TABLE IV. RESULTS OF CNN+GloVe, LSTM CNN+GloVe, CNN-LSTM+GloVe, AND LSTM-CNN+GloVe

Label	CNN+GloVe						LSTM+GloVe						LSTM-CNN+GloVe						LSTM-CNN+GloVe					
	P	R	F1	S	Loss	Acc	P	R	F	S	Loss	Acc	P	R	F1	S	Loss	Acc	P	R	F1	S	Loss	Acc
Domestic Violence	0.99	0.99	0.99	3577	0.04	0.98	0.98	0.99	0.98	3577	0.01	0.98	0.99	0.99	0.99	3577	0.03	0.98	0.98	0.98	0.98	3577	0.12	0.97
Femicide	0.62	0.63	0.62	235			0.72	0.68	0.70	235			0.60	0.73	0.66	235			0.62	0.63	0.63	235		
Forced Abortion	0.96	0.96	0.96	285			0.99	0.97	0.98	285			1.00	0.98	0.99	285			0.98	0.93	0.96	285		
Forced Marriage	0.97	0.98	0.98	584			0.99	0.99	0.99	584			0.98	0.99	0.99	584			0.98	0.98	0.98	584		
Online Violence	0.91	0.89	0.90	253			0.99	0.95	0.97	253			0.97	0.94	0.95	253			0.97	0.90	0.93	253		
Rape	0.98	0.98	0.98	5439			0.99	0.99	0.99	5439			0.99	0.99	0.99	5439			0.98	0.98	0.98	5439		
Sex Trafficking	0.99	0.98	0.98	1369			0.99	0.98	0.99	1369			0.99	0.98	0.99	1369			0.99	0.97	0.98	1369		
Sexual Harassment	0.98	0.98	0.98	5041			0.99	0.99	0.99	5041			0.99	0.98	0.98	5041			0.97	0.98	0.98	5041		
Average	0.93	0.92	0.92	1678			0.95	0.94	0.95	1678			0.94	0.95	0.94	1678			0.93	0.92	0.93	1678		

TABLE V. COMPUTATIONAL TIME DURING TRAINING

Model	Acc	Loss	Computational Time
CNN	1.00	0.000000042122	1 h 27 min 13 s
LSTM	0.99	0.00049133	1 h 40 min 17 s
CNN-LSTM	0.99	0.00060757	3 h 37 min 11s
LSTM-CNN	1.00	0.000000054554	1 h 59min 27 s
CNN+GloVe	1.00	0.000046438	9 min 10 s
LSTM+GloVe	0.99	0.0064	44 min 28 s
CNN-LSTM+GloVe	0.99	0.0067	22 min 35s
LSTM-CNN+GloVe	0.99	0.00069372	33 min 37 s

## V. DISCUSSIONS

This study finds that the LSTM model using the GloVe word embedding pre-train model delivers the best results after extensive training and testing. To classify the model's output, the following parameters were used: a 100-layer LSTM hidden layer, a max pooling layer, a relu activation function used at the first dense layer and a softmax activation function on the second dense, a learning rate of 0.0003 with the Adam optimizer, and a total of 20 epochs. Metrics such as the gap between the two sets, the accuracy of both sets, and the precision, recall, and f1-score value reveal differences between the training and testing sets.

The gap between the two measures of accuracy, training, and validation, narrows to a reasonable level during model training. In comparison, other models' validation accuracy becomes linear after a few epochs, although training accuracy is substantially higher. Another model has been overfitted, but because the FP is the measure, the LSTM has very little overfitting. Furthermore, during the testing phase, the LSTM with the GloVe embedding word had the maximum consistency across all three performance parameters. It is supported by the capability offered by GloVe [25].

Furthermore, the data with the fewest labels has the lowest precision, recall, and f1 score. Throughout the experiment, the femicide-labeled data has the lowest precision, recall, and f1-score. Most labels usually result in the best accuracy, recall, and f1-score.

All models with and without GloVe test findings are elaborated. Deep learning models can effectively categorize tagged text without using a pre-trained GloVe. The accuracy of CNN is 0.982, followed by LSTM-CNN of about 0.981, and then LSTM is 0.980. Although the dataset was unbalanced, the model nevertheless achieved respectable levels of accuracy. The outcomes ranged from 0.94 to 0.96. According to the study's preliminary settings, all GloVe-based models perform admirably on the testing set. Tagging is not required to succeed in a deep learning system that does not use a pre-trained GloVe as a word embedding model. In terms of accuracy, LSTM+GloVe is superior to CNN+GloVe, CNN-LSTM+GloVe, and LSTM-CNN+GloVe at 0.983. The model's accuracy, recall, and f1-score are all within an acceptable range 0.92 to 0.95 despite using an unbalanced dataset.

Furthermore, when comparing standard word embedding to GloVe's pre-trained word embedding, deep learning models using GloVe show significant improvement, particularly in computing time. When GloVe word embedding is not utilized, the computational time for all three models combined exceeds an hour: 1 hour 27 minutes for the CNN model, 1 hour 40 minutes for the LSTM model, and 1 hour 59 minutes for the LSTM-CNN model. When employing GloVe word embeddings, the CNN model takes 9 minutes, the LSTM model 44 minutes, and the LSTM-CNN model 33 minutes to compute.

All models that classify femicide class have produced the lowest result in precision, recall, and f1-score. This could be because the femicide class has the lowest data among all the classes. Meanwhile, the largest class, such as domestic

violence, sexual harassment, and rape, tend to have the highest precision, recall, and f1-score. More research on multi-class text classification is required to obtain a better result [26].

## VI. CONCLUSIONS

This research compares machine learning models that utilize the GloVe and without GloVe embedding methods to see if there is an improvement in text multi-classification problem-solving. The proposed hybrid LSTM-CNN and CNN-LSTM with GloVe and without GloVe can classify the multi-class text. However, the experimental results prove that the effectiveness, capability, and efficiency of the LSTM-CNN and CNN-LSTM with GloVe significantly improved the multi-class performance in GV tweet data compared to those without GloVe. It is also better than a single CNN and LSTM in terms of accuracy. It can be said that the hybrid solution and embedded GloVe have demonstrated a reduction in computational time. Thus, it is expected that the hybrid LSTM-CNN and CNN-LSTM with GloVe can be used in other domains. In the future, evaluating the tweet text data from a different domain and considering larger multi-class datasets are recommended.

## ACKNOWLEDGMENT

The authors would like to acknowledge Institute for Big Data Analytics and Artificial Intelligence (IBDAAI), Universiti Teknologi MARA, for the financial support provided to this research project.

## REFERENCES

- [1] M. Castorena, I. M. Abundez, R. Alejo, E. E. Granda-Gutiérrez, E. Rendón, and O. Villegas, "Deep Neural Network for Gender-Based Violence Detection on Twitter Messages," *Mathematics*, vol. 9, no. 8, 2021. doi: 10.3390/math9080807.
- [2] R. Capucci, C. Paganelli, S. Carboni, R. Cappadona, M. Roberto, and G. Rinaldi, "Characteristics of Gender-Based Violence Determined from Emergency Room Visits," *Violence Gen.*, vol. 2, no. 2, pp. 129–133, Jun. 2015, DOI: 10.1089/vio.2014.0034.
- [3] M. Mohan, "One in three women are subjected to violence - WHO," *BBC News*, 2021. <https://www.bbc.com/news/world-56337819> (accessed December 19, 2021).
- [4] J. A. Odera and J. Mulusa, "SDGs, gender equality and women's empowerment: what prospects for delivery?" *Sustainable development goals and human rights*: Springer, pp. 95–118, 2020.
- [5] E. Chuck, "#MeToo: Alyssa Milano promotes hashtag that becomes anti-harassment rallying cry," *NBC News*, 2017. <https://www.nbcnews.com/storyline/sexual-misconduct/metoo-hashtag-becomes-anti-sexual-harassment-assault-rallying-cry-n810986> (accessed January 05, 2022).
- [6] F. Hanafi, "Watch: Female Passenger Gets Harassed By E-Hailing Driver," *World of Buzz*, 2022. [https://worldofbuzz.com/watch-female-passenger-get-sexually-harassed-by-her-e-hailing-driver/?fbclid=IwAR3o2sM7e6w4Ot\\_4IRCpNhVUhtnRlMnHCHmaIlrzTMFh86Ob30bHeVYNdE](https://worldofbuzz.com/watch-female-passenger-get-sexually-harassed-by-her-e-hailing-driver/?fbclid=IwAR3o2sM7e6w4Ot_4IRCpNhVUhtnRlMnHCHmaIlrzTMFh86Ob30bHeVYNdE) (accessed April 14, 2022).
- [7] J. A. García-Díaz, M. Cánovas-García, R. Colomo-Palacios, and R. Valencia-García, "Detecting misogyny in Spanish tweets. An approach based on linguistics features and word embeddings," *Futur. Gener. Comput. Syst.*, vol. 114, pp. 506–518, 2021, DOI: 10.1016/j.future.2020.08.032.
- [8] N. Suzor, M. Dragiewicz, B. Harris, R. Gillett, J. Burgess, and T. Van Geelen, "Human rights by design: The responsibilities of social media platforms to address gender - based violence online," *Policy & Internet*, vol. 11, no. 1, pp. 84 – 103, 2019.
- [9] A. Meco, Lucina De, & Mackay, "Social media, violence and gender norms: The need for a new digital social contract," *Align Platform*,

2022. <https://www.alignplatform.org/resources/blog/social-media-violence-and-gender-norms-need-new-digital-social-contract> (accessed April 15, 2022).
- [10] A. Sahay, "The silenced women: What works in encouraging women to report cases of gender-based violence?" World Bank Blogs, 2021. <https://blogs.worldbank.org/developmenttalk/silenced-women-what-works-encouraging-women-report-cases-gender-based-violence> (accessed April 15, 2021).
- [11] S. Mittal and T. Singh, "Gender-based violence during COVID-19 pandemic: a mini-review," *Front. Glob. women's Heal.*, p. 4, 2020.
- [12] A. Khatua, E. Cambria, and A. Khatua, "Sounds of silence breakers: Exploring sexual violence on twitter," in 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), 2018, pp. 397–400.
- [13] E. Alawneh, M. Al-Fawa'reh, M. T. Jafar, and M. A. Fayoumi, "Sentiment Analysis-Based Sexual Harassment Detection Using Machine Learning Techniques," in 2021 International Symposium on Electronics and Smart Devices (ISESD), 2021, pp. 1–6. DOI: 10.1109/ISESD53023.2021.9501725.
- [14] M. Zakkar and D. Lizotte, "Analyzing Patient Stories on Social Media Using Text Analytics," *J. Healthc. Informatics Res.*, vol. 5, Dec. 2021, DOI: 10.1007/s41666-021-00097-5.
- [15] H. ALSaif and T. Alotaibi, "Arabic text classification using feature-reduction techniques for detecting violence on social media," *Int. J. Adv. Comput. Sci. Appl.*, vol. 10, no. 4, 2019.
- [16] L. M. Both, L. Helena, M. Freitas, and I. Passos, "Study using Text Classification Tools," vol. 20, no. 2, 2020.
- [17] I. Soldevilla and N. Flores, "Natural Language Processing through BERT for Identifying Gender-Based Violence Messages on Social Media," in 2021 IEEE International Conference on Information Communication and Software Engineering (ICICSE), 2021, pp. 204–208. DOI: 10.1109/ICICSE52190.2021.9404127.
- [18] M. B. Mutanga and A. Abayomi, "Tweeting on COVID-19 pandemic in South Africa: LDA-based topic modelling approach," *African J. Sci. Technol. Innov. Dev.*, vol. 14, no. 1, pp. 163–172, 2022.
- [19] D. Ofer, "Machine Learning for Protein Function," Mar. 2016.
- [20] U. N. in Iran, "Frequently asked questions: Types of violence against women and girls," I. R. Iran, 2020. <https://iran.un.org/en/102394-frequently-asked-questions-types-violence-against-women-and-girls> (accessed May 05, 2022).
- [21] GBVIMS, "The Gender - Based Violence Classification Tool." gender-based violence information management system, 2021.
- [22] L. Hickman, S. Thapa, L. Tay, M. Cao, and P. Srinivasan, "Text preprocessing for text mining in organizational research: Review and recommendations," *Organ. Res. Methods*, vol. 25, no. 1, pp. 114–146, 2022.
- [23] S. U. Hassan, J. Ahamed, and K. Ahmad, "Analytics of machine learning-based algorithms for text classification," *Sustain. Oper. Comput.*, vol. 3, pp. 238–248, 2022, DOI: <https://doi.org/10.1016/j.susoc.2022.03.001>.
- [24] I. Markoulidakis, I. Rallis, I. Georgoulas, G. Kopsiaftis, A. Doulamis, and N. Doulamis, "Multi-class Confusion Matrix Reduction Method and Its Application on Net Promoter Score Classification Problem," *Technologies*, vol. 9, no. 4, 2021, DOI: 10.3390/technologies9040081.Kowsari.
- [25] K. Meimandi, K. J. Heidarysafa, M. Mendu, S. Barnes, L. and D. Brown. "Text classification algorithms: A survey Information (Switzerland)," vol 10, no. 4, pp. 1–68, 2019, <https://doi.org/10.3390/info10040150>.
- [26] Y. Arslan, K. Allix, L. Veiber, C. Lothritz, T. F. Bissyandé, J. Klein, and A. Goujon, "A comparison of pre-trained language models for multi-class text classification in the financial domain," In Companion Proceedings of the Web Conference, pp. 260–268, 2021.



# Performance Analysis of Deep Learning YOLO Models for South Asian Regional Vehicle Recognition

Minar Mahmud Rafi, Siddharth Chakma, Asif Mahmud, Raj Xavier Rozario, Rukon Uddin Munna, Md. Abrar Abedin Wohra, Rakibul Haque Joy, Khan Raqib Mahmud, Bijan Paul\*

Department of Computer Science and Engineering  
University of Liberal Arts Bangladesh  
Dhaka, Bangladesh

**Abstract**—For years, humans have pondered the possibility of combining human and machine intelligence. The purpose of this research is to recognize vehicles from media and while there are multiple models associated with this, models that can detect vehicles commonly used in developing countries like Bangladesh, India, etc. are scarce. Our focus was to assimilate the largest dataset of vehicles exclusive to South Asia in addition to the more common universal vehicles and apply it to track and recognize these vehicles, even in motion. To develop this, we increased the class variations and quantity of the data and used multiple variations of the YOLOv5 model. We trained different versions of the model with our dataset to properly measure the degree of accuracy between the models in detecting the more unique vehicles. If vehicle detection and tracking are adopted and implemented in live traffic camera feeds, the information can be used to create smart traffic systems that can regulate congestion and routing by identifying and separating fast and slow-moving vehicles on the road. The comparison between the three different YOLOv5 models led to an analysis that indicates that the large variant of the YOLOv5 architecture outperforms the rest.

**Keywords**—*You Only Look Once (YOLOv5); vehicle detection; neural network; deep learning; vehicle tracking*

## I. INTRODUCTION

Advancements in automobile manufacturing have given rise to more affordable cars which has resulted in over five million registered vehicles [1] coasting through the roads of Bangladesh. Road infrastructures in this country were not designed to hold the growing number of vehicles which presents grave environmental and health concerns. Given the circumstance, congestion is inevitable, and this significantly contributes to the rising air and noise pollution levels in the city. To circumvent this obstacle, restless drivers resort to maneuvering chaotically without any regard to traffic rules and are thus responsible for most of the road fatality cases in the country. One of the biggest obstacles is that traditional methods of prevention such as traffic lights and pedestrian crossings are not sufficient because they are generally ignored.

To address this issue, an intuitive system is needed to observe traffic patterns and direct different vehicles into proper lanes. Most vehicles in South Asia are very different than those in the western world as they differ drastically in shapes, sizes, and colors. This is a major challenge the algorithm will face [2]

as it needs to differentiate between these vehicles to identify them individually. The height and angle at which these vehicles are posed and captured also factor into this problem. Datasets that include traditionally used South Asian vehicles are scarce and do not contain the required amount of data which presents a separate challenge. Due to the erratic nature of traffic in South Asian countries, different CNN models that are usually tested in other environments have not been applied enough to see how they perform in the tumultuous streets of cities like Dhaka. A major challenge of our research is that we have had data scarcity, particularly for south Asian vehicles.

Machine learning has progressed enough to make use of traffic cameras [3] to track vehicles and their patterns. Additionally, using Neural Network-based Object Detection can produce valuable tracking and surveillance data that could be essential to coming up with a solution to the traffic problem. Further applications in the division of slow- and fast-moving vehicles and the identification of missing vehicles can also be pursued through Deep Learning. Smart traffic systems [4] can utilize these applications to reduce mishaps while also improving the flow of traffic. Autonomously driven cars [5] can also employ the previously mentioned applications to avoid different vehicles, clogged roads, and potential accidents while on the road.

However, one of the key difficulties in using machine learning algorithms is the requirement of a vast amount of data to train a model. In this research, we develop a sizable vehicle dataset from scratch and train a model to accurately recognize them. The intention was to set our work apart from conventional vehicle detection systems. Our research is distinctive in that we have curated a dataset consisting of 21 classes of vehicles commonly available worldwide and those that are only seen in South Asian regions. Unique vehicles like rickshaws, human haulers, three-wheelers, etc. all vary in build and proportion. The collected images are put through a lengthy process of cleaning, augmenting, and finally labeling through bounding box annotations. To address the data scarcity issue, we used different augmentation techniques to balance the dataset. This is done to ensure we have enough data for accurate testing and training. We chose a well-known object detection algorithm called YOLOv5 (You Only Look Once) [6] to use in our model for training and we compared the

\*Corresponding Author.

performance of different architectures of YOLOv5 models: Small, Medium, and Large. Previous versions [7] were an option but the new update presented a more efficient and time-saving alternative. YOLOv5's hyperparameters are tweaked to accurately detect objects in real-time through bounding box coordinates of objects from the carefully labeled data it has been trained with. This model will be able to recognize different native vehicles which can then be used in systems to reduce the traffic problem that plagues South Asian cities.

## II. LITERATURE REVIEW

The subject of object detection has attracted attention from various independent researchers over the years. Different detection systems were used over the years for identifying objects. LIDAR (Light Detection and Ranging) was used in the form of sensors attached to both vehicles and certain points of the road to detect oncoming vehicles [8]. Other non-intrusive methods like ASFF (Adaptive Spatial Feature Fusion) [9] and Radar Sensors [10] were also used. The interest in this field and its innovations date back to the 1970s [11].

Object detection [12] by the camera has become more prevalent in recent years and more accurate and cost-effective than other sensors. To accurately identify vehicles, real-time detection speed and high accuracy are required for a quick response to fast-moving vehicles and to get a reduced latency. While most algorithms repurpose classifiers by taking images at multiple scales and locations and applying the algorithm to perform detection [13], YOLO applies a neural network that dissects an image into different parts and can predict bounding box regions based on predicted probabilities [14]. YOLO detects objects using a single inference which makes it faster than its peers, SSD (Single Shot Detector) and Faster R-CNN (Region-based Convolutional Neural Network) [15].

Research by Liu & Zhang [16] aimed to improve the standard YOLOv3 model by training it to adapt to actual traffic conditions and applying a scale prediction layer to improve the detection accuracy of large vehicles. They use the k-means++ algorithm to improve the efficiency of the anchor box dimension clustering as shown in Fig. 1. The resulting F-YOLOv3 algorithm clocked in at 91.12% on the mAP (Mean Average Precision) accuracy scale beating out Faster R-CNN at 90.01% and base YOLOv3 at 78.68%. The recognition performance of large vehicles is poor when compared to small vehicles because of their contrasting characteristics.

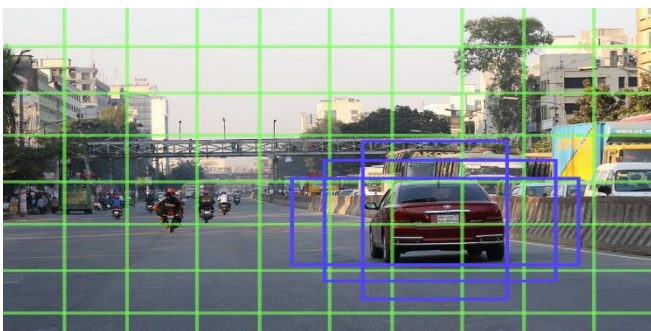


Fig. 1. Anchor Box Dimension Clustering [16].

Redmon [17] suggested the integration of classification and localization into a single convolutional neural network which would improve the speed at the cost of precision. While it achieves a combined accuracy of 75.0%, the model has difficulty in detecting smaller grouped objects due to the spatial constraints imposed by YOLO and objects in different aspect ratios. Chandan suggested a different approach [18] where he opted to use OpenCV to detect objects in a python environment with the assistance of the Single Shot Detector algorithm. This algorithm used optical flow and background subtraction to achieve an optimal accuracy in detecting standard vehicle classes and this was a great basis for comparison with the earlier versions of YOLO. A detection system for localized mobile environments like roads and railways was made by Chen [19]. Using the same COCO dataset, they compared YOLOv3, and the Single Shot Detector mentioned above to find their efficiency and applicability in traffic. It was found that YOLOv3 had attained an 85% score over SSD's 79.5% in terms of mAP at high resolution. A more recent comparison of YOLOv4 with SSD and Faster R-CNN was conducted by Kim [20] for real-time vehicle detection. After evaluating the different models, it was observed that YOLOv4 performed at 98.1% precision while Faster R-CNN and SSD performed at 93.4% and 90.5% respectively.

Phillips suggested a system [21] for distance estimation between vehicles in traffic to avoid collision by mounting a monocular camera to a vehicle dashboard. It is fitted with systems for object tracking and detection and is modular enough to switch out systems for other uses. Errors in estimation increase as the distance increases. Wang has used edge Detection technology [22] to demonstrate the detection of objects such as vehicles by their outer edge lines. The edge detection technique must remove noises from an image background using a higher threshold before identification of the vehicle in question can begin. This changes how we can detect vehicles from a certain height and makes detection possible using image-capturing objects like Drones.

Sokalski [23] produced another alternative that combines edge detection with color identification to differentiate between artificial and natural objects. The only drawback is the process of extracting the nine features from various channels of each color in the image which are used to define the edges. An identification approach by thickness estimation and edge detection was put forward by Kanistras [24] where angle vectors of an elevated image would be determined in its edge guide. These vectors are constantly changing by determining the standard deviation of slope vectors therefore pre-defining edges to detect vehicles.

Different datasets and their use in creating a large diverse dataset in the training of algorithms are discussed by Xiao & Kang [25]. This paper has influenced how we approached diversifying and enriching our dataset to obtain satisfactory test results. The paper also provides tips for being efficient in collecting and labeling datasets properly. The importance of data augmentation is made clear by Zoph [26] in his discussions about how different augmentation strategies such as rotating, shearing, equalizing, changing colors, etc. can not only expand the dataset but also increase accuracy up to 6% but at the cost of data loss during training.

A YOLO-based traffic counting system developed by Lin [27] was employing three different pieces. The detector generates the bounding boxes of the vehicles, the buffer stores the vehicle coordinates, and the counter is responsible for counting the vehicles. Images/videos are put through the detector where it passes through filters and then the YOLO algorithm. Data access is built from the frame number input and output, previous and current array in buffer, and vehicle counting algorithm in counter. Checkpoints are also added for validation of detection whereas the overall accuracy is determined by using a video that has a different height and angular perspective. The counting accuracy seen during the day was around 95% but dropped to unfavorable rates at night due to factors such as headlight exposure, dim streetlights, etc. which was later improved upon by implementing night vision technology. An alternate YOLO method created by Tao [28] removes the last two layers of the connected system and adds a pooling layer. This is faster than its peers and the addition of a pre-processing procedure for night images enhances detection in darkness by removing highlights and modifying contrast and brightness. This new optimized O-YOLO algorithm executes an accuracy of 66% on a standard VOC dataset and 80.1% on a custom-curated dataset. Corovic [29] implemented YOLO to detect objects in real-time traffic and pre-trained the algorithm to detect them in five categories. These were cars, trucks, civilians, signs, and lights. Tests were conducted to prove that YOLO was suitable for real-time detection and in different weather. They deducted detection could be improved in place of obstructions by incorporating datasets that contain weather conditions into the training. After training of 120 epochs was conducted, the accuracy had a steady increase from 18.98% to 46.60% but failed to climb to higher rates due to many occluded objects in the dataset. Salarpour [30] realized an algorithm to track multiple vehicles using the Kalman filter and background subtraction. A region-based algorithm is then combined with the filter to track and predict the region of the vehicle in the continuing frame while also using its color and size to get an accurate result of 96%. This method helps detect issues such as occlusion and clutter with minimal loss of accuracy.

Occlusion makes it hard for vehicles to be distinguished for detection and therefore a procedure to reduce dense occlusion from surveillance cameras was put forward by Phan [31]. This is also a combination of background subtraction and detection but mixed with occlusion detection where each occluded vehicle is extracted from images based on their features. The method improves the accuracy of detection in occluded images at higher angles proven by its 85% accuracy score during high traffic. To address the problem of detecting vehicles at varying scales and distances, Lu [32] produced a modified version of the Region Proposal Network (RPN) which is tailored to be scale aware during detection. This system has two different sub-networks to detect large and small case proposals and inputs through two separate XGBoost (Extreme Gradient Boosting) algorithms to create final predictions. Both algorithms boasted scores of 64.1% and 84.8% respectively in terms of precision.

The most consistently accurate model out of the many commented on above is the YOLOv4 algorithm. It offers a

staggering 98.1% mean average precision when applied over a vehicle recognition system. YOLOv5, which currently lacks substantial research documentation, has data that exhibits improved accuracy and speeds over its previous iteration [33] which will be nothing but beneficial to our training.

To train a varied dataset such as this, we needed a sustainable system powerful enough to process and execute all the data. The model we used in conjunction with neural networks was built to enhance every aspect of its previous build and therefore the database was processed much quicker than expected. The database itself is a mixture of both common and unique vehicles found here in the South of Asia but the rarity in variety of some of these vehicles was a complication. 21 assorted classes were selected, collected, processed, and augmented to create a robust dataset for this research.

### III. ARCHITECTURE

The method of detecting items in an image as shown in Fig. 2 and calculating their location using bounding boxes is known as object detection. The classification of images is concerned with determining whether an object exists in each image based on calculative likelihood. Images have characteristics like distinct edges that an object recognition method must extract. Convolutional Neural Networks, Auto Encoder techniques, and others, can be used to automate this procedure. The most effective object identification strategy is one that assures that all objects of vivid size are given a bounding box to be recognized, as well as having high computational capabilities allowing for faster processing. Both YOLO and SSD promise good outcomes, but there is a speed/accuracy trade-off.

#### A. Proposed Methodology

In the South Asian region, we began gathering images of various vehicle items. The dataset was sorted and categorized after the image collections were completed to prepare it for the machine learning model. This dataset was then split at random using the standard splitting technique, with 80% of the data going to the training set for training and 20% going to the validation set for validation.



Fig. 2. Detection Process during a Training Phase.



This is done to assess the model's correctness while avoiding over-fitting. The prediction model was evaluated using the set developed for the validation procedure, whereas the training set provided the algorithm. We were given numerous metrics and statistics to evaluate because of the outcome. The whole process is outlined in Fig. 3.

The YOLO model is part of the Fully Convolutional Neural Networks (FCN) family that allows for the most optimally achievable outcome and real-time object recognition in every end-to-end model. YOLOv5 as shown in Fig. 4 contains a variety of internal models, each with its own set of complexity and architectures. The largest and most sophisticated network is YOLOv5l, which is followed by YOLOv5m and YOLOv5s.

Each of these models was trained to see how the various architectures influence the overall model's speed and accuracy. The three different and crucial aspects of YOLOv5's architecture can be outlined.

**Backbone** - The backbone is primarily responsible for resolving gradient information and incorporating changes into feature maps, hence lowering parameter numbers and the overall model's FLOPS (Floating Point Operation per Second).

**Neck** - The neck improves the data flow within the model. It includes a feature pyramid network with an upgraded bottom-up approach that can expose new low-level complicated features, as well as localization signals in lower layers that can improve localization accuracy. Through interconnection provided by Adaptive Feature Pooling, the feature grids, and levels produce useful data on all levels.

**Head** - The head contains algorithms for creating feature maps of many sizes with prediction procedures that the model employs to recognize objects of many sizes. Because vehicles might have components and add-ons of various shapes and sizes, this technique is critical for vehicle recognition. Each of these items can be easily detected using the multi-scale detection feature. When the training on the model is started in YOLOv5, the procedure begins.

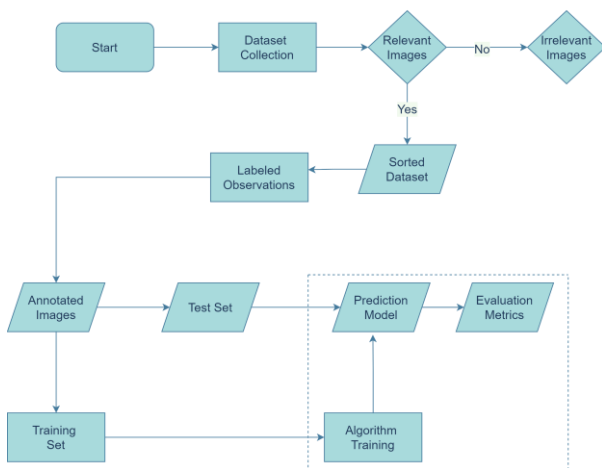


Fig. 3. Block Diagram of Vehicle Detection & Tracking Process.

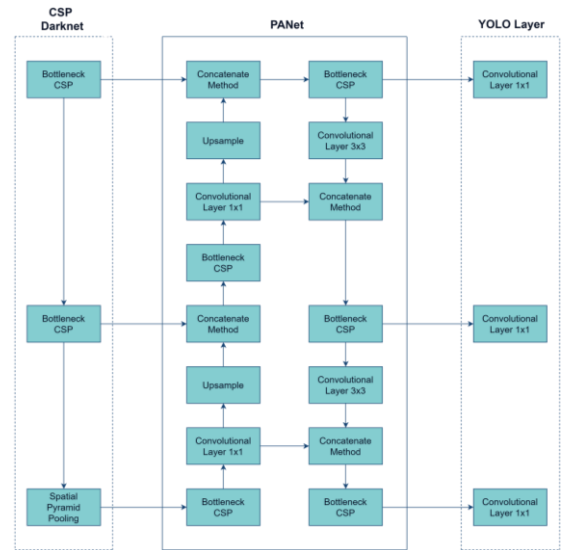


Fig. 4. YOLOv5 Architecture

The data is fed into the Sparked module, which extracts the features, which are subsequently transmitted to the PANet (Path Aggregation Network) [34] module to be fused.

It is all gathered in the YOLO layer, which is then processed to produce important analytic data like class, location, score, and size.

#### B. Data Annotation Format in YOLOv5

Each image was annotated in the form of an a.txt file, with each line of the content record depicting a bounding box. In Fig. 5, for example, there are four items (car, minivan, pickup, and bus).

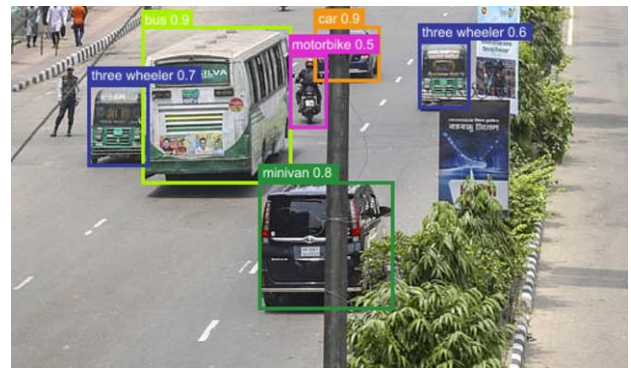


Fig. 5. Bounding Boxes For Data Annotation.

### IV. DESIGN AND IMPLEMENTATION

On the vehicle dataset, three distinct YOLOv5 architectures – small, medium, and large were implemented and trained. Table I lists the dataset metaphors that have been defined. The total number of images and their respective classes are displayed in Fig. 7.

This dataset needed to be carefully sliced for training. There was no exact way to divide the dataset, so we employed the traditional slicing technique. We separated 80% of the data for training and a small 20% for the validation process which is required to reduce over-fitting. A random selection was made

from the main dataset to create the test set which was sufficient for the calculation of the accuracy of the model.

The final dataset contains 11,808 images of different classes of vehicles, out of which 9,749 images were used for training and 2059 images were used for validation. After splitting the dataset and annotating the images, the training set and validation set were fed into the machine learning model. The training set was used for the algorithm training process and the validation set was used for the prediction model which provided us with the different evaluation metrics and statistics.

The instances of some of the classes are higher than others and this makes the dataset non-uniform in nature. For example, the 'Car' class contains the highest number of appearances in our dataset with a total number of 10,680 whereas classes like 'garbage van' and 'Police car' appear less in the dataset.

**A. Dataset Preparation**

We used a custom data collection with about 21 types of automobiles in South Asian territory for our research. Most of the images were gathered from real-time data acquired by users, social networking pages, blogs, and other online sources, and the improper images were filtered out of the dataset.

Filtering the data collection is impressive, and as with any model preparation, it is required to increase the amount of relevant data that our model can extract from the dataset. Because there are images containing items that aren't supposed to be there, the dataset is full of noise.

**B. Dataset Pre-processing and Augmentation**

When it comes to improving the model's performance, pre-processing the dataset is essential. It is a necessary step toward improving the quality of data and the amount of useful information the model can derive from it. It is also critical to generate a balanced dataset to improve the accuracy of an

existing model. Before the dataset was supplied to the model for training, the images with their pixels had to be reshaped and resized. The images were resized using the `numpy.reshape()` method, and the pixels were replaced with `pixel/255` using vector scalar division.

Fig. 8 shows how the different augmentations were carried out and as a result, multiple different versions of the image were created. Balancing this dataset was a complex task as different categories of images had massive disparities in numbers.



Fig. 6. Sample Images from the Dataset.

TABLE I. DATASET SPECIFICATIONS

Attributes	Features
Image Type	RGB
Image Extension	JPG, PNG
Image Dimension	1920 * 1920

**Dataset Visualization After Augmentation**

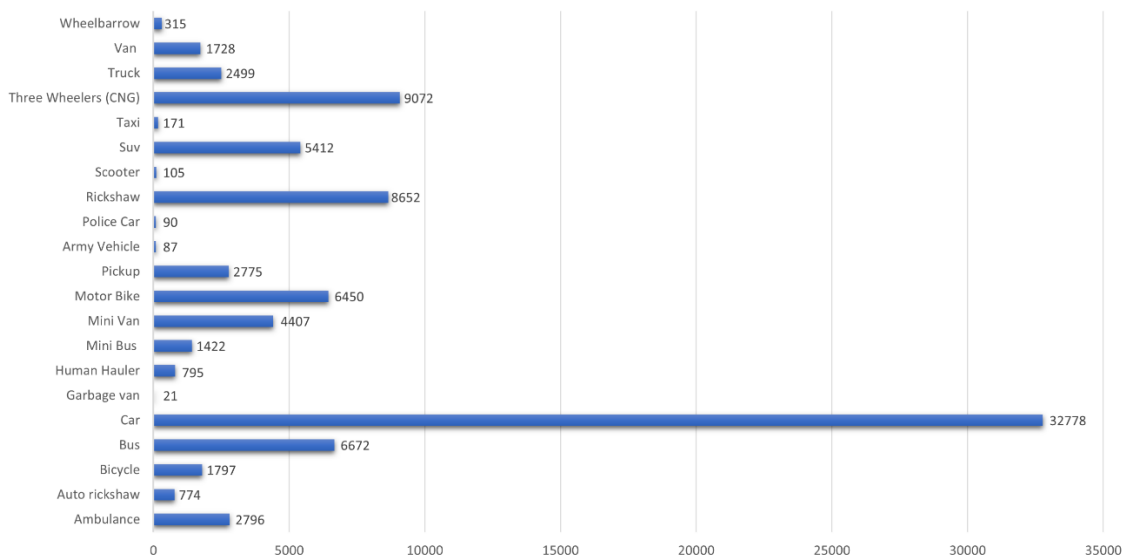


Fig. 7. Number of Instances Per Class before Augmentation.



Fig. 8. Different Types of Image Augmentations Applied to Our Dataset.

To solve this, only two different augmentations were carried out. The Contrast and Grayscale filters were utilized on our dataset to create augmented data with distinct differences that would be easily identifiable by the machine when trained. Because our original dataset was imbalanced and non-uniform, augmentation was used to increase the number of instances of classes in the dataset, particularly those that appeared less frequently like ambulances, bicycles, etc., thus improving the model's accuracy and performance by making the dataset more balanced.

After augmentation, a relative balance within the dataset was achieved which can be seen in Fig. 9 which shows the instances of different classes after augmentation was applied to it compared to Fig. 6 above which was before augmentation.

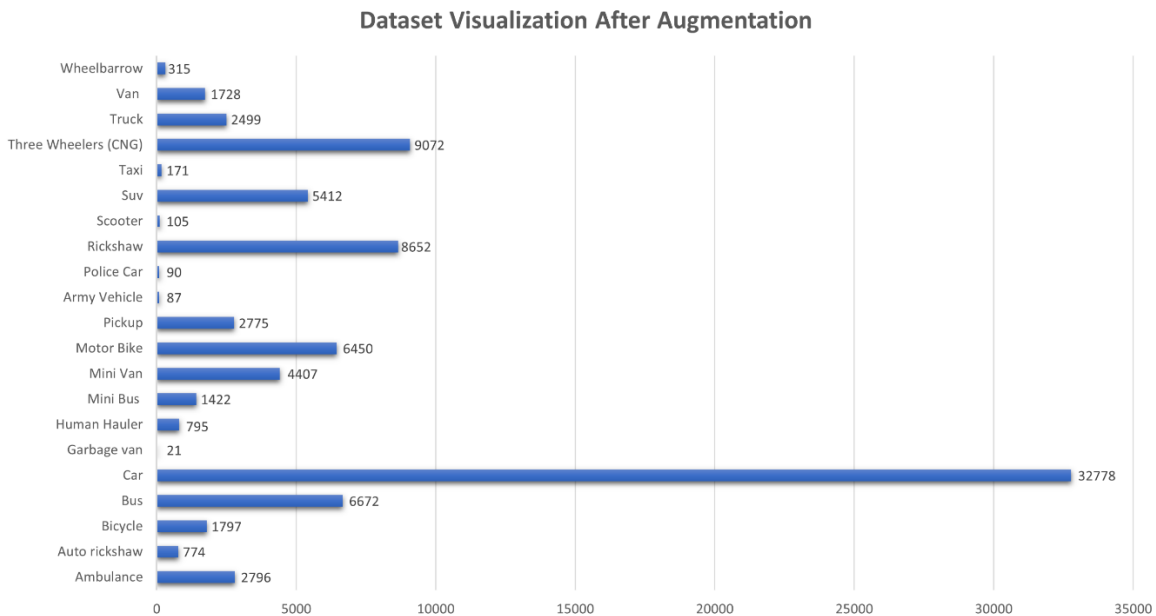


Fig. 9. Number of Instances Per Class after Augmentation.

### C. Model Training

YOLOv5 has multiple architectures such as YOLOv5s, YOLOv5m, and YOLOv5l in order of complexity and depth. We have implemented and juxtaposed each of these models on our dataset to determine the one which fits best. Each of the models was trained for 100 epochs which took around 12 hours per model. The model training process is illustrated in Fig. 10. The network is trained using a collection of training data, and it then learns to predict the target values. We have also improved the accuracy of the dataset for the YOLOv5l, YOLOv5m, and YOLOv5s designs.

An appropriate dataset is required to train a deep learning network. You may need to make a Train-Test Split depending on the available data. During the training phase, validation losses are tracked, and non-constant values are generated after several epochs. Otherwise, the model will be adjusted for hyper-parameters, and the validation loss value will be kept as low as possible.

The model with the largest validation loss is saved for testing on the real data. When a model obtains high precision and recall rates for new datasets, or when it demonstrates improved performance after training on an enriched dataset, it is said to perform satisfactorily.

### D. Dataset Demonstration

For model training, the entire vehicle dataset is provided. Table II shows the number of images evaluated for implementation, the number of images used to train the model, and the number of images used in testing.



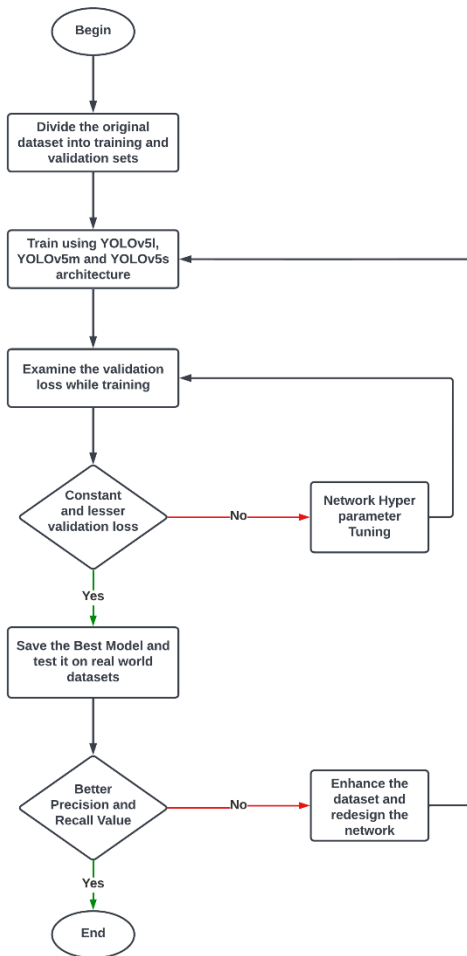


Fig. 10. Flow Diagram of Model Training.

TABLE II. DATASET DISTRIBUTION

Features	Before Augmentation	After Augmentation
Number of Classifiers	21	21
Total Input Images	11,808	88,818
Images to be Trained	9,749	55,956
Images to be Tested	2,059	32,862

## V. DISCUSSION

We were able to create a big dataset with over 21 unique classes and about 11,808 images. The dataset size increased to exactly 88,818 instances after the various Data Augmentation methods were applied.

### A. Performance Matrices

**Confusion Matrix:** Assists us in this by providing a comprehensive assessment of each model's performance, including faults.

- TP (True Positive) - When the anticipated value is equal to the actual value and the result is positive.
- TN (True Negative) - When the projected value is the same as the actual value, but the value is negative.

- FP (False Positive) - The anticipated value was incorrectly predicted as positive when the actual value was negative.
- FN (False Negative) - The anticipated value was incorrectly predicted as negative when the actual value was positive.

A variety of performance measures can be calculated using these numbers.

**Accuracy:** Measurement as a singular metric is invalid because it assigns equal costs to different types of errors and can only be used with a well-balanced dataset. The formula is as follows:

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN} \quad (1)$$

**Precision:** Precision is a metric for how many accurately anticipated situations turn out to be positive, which can help establish a model's eligibility.

The formula for precision is as follows:

$$\text{Precision} = \frac{TP}{TP+FP} \quad (2)$$

**Recall:** Recall is the measure of the positive cases that were correctly classified by our model and is defined by the following formula:

$$\text{Recall} = \frac{TP}{TP+FN} \quad (3)$$

**mAP (mean Average Precision):** mAP is used to evaluate object detection models like Fast R-CNN, YOLO, and Mask R-CNN. It considers both types of errors, false positives (FP) and false negatives (FN), as well as the trade-off between precision and recall (FN).

$$\text{mAP} = \frac{1}{N} \sum_{i=1}^N AP_i \quad (4)$$

The mAP is calculated by averaging each class's Average Precision (AP) over several classes.

**F1 Score:** It combines the values of Precision and Recall into a single metric that must be maximized to enhance our model. However, interpreting the F1 score is challenging, leaving us oblivious to which of the metrics the model is optimizing. The formula is as follows:

$$\text{F1 Score} = \frac{2}{\frac{1}{\text{Recall}} + \frac{1}{\text{Precision}}} \quad (5)$$

**Loss function:** We will summarize YOLOv5 losses and metrics to help you better comprehend the outcomes. The three parts of the YOLO loss function are box loss, obj loss, and class loss.

## VI. RESULT ANALYSIS

For this research, three distinct YOLOv5 designs – small, medium, and large – were implemented and used. Before training, the dataset was pre-processed and supplemented. The complexity and depth of the three YOLOv5 architectures differ. Each model was tested against our dataset, and the results were compared to determine which model performed the best in terms of vehicle detection.

Mean average precision is a well-known and commonly used object detection evaluation metric. Faster R-CNN [35], YOLO [36], and MobileNet [37] are all state-of-the-art models that use mAP to evaluate their models. We tried to test the performance of the three YOLOv5 models – small, medium, and large – in our implementation.

### A. Mean Average Precision (mAP) Analysis

The mAP 0.5 of the three designs employed throughout 100 epochs is illustrated in Fig. 11 and Fig. 12. The YOLOv5l model achieves more accuracy in Fig. 11 and Fig. 12 than the other two models both before and after augmentation, as shown in the graphs.

### B. Training Loss Analysis

The training loss is a metric that measures how well a deep learning model matches the training data. That is, it evaluates the model's error on the training data. The training set is a subset of the dataset that was used to train the model originally. The training loss is calculated computationally by adding the sum of errors for each sample in the training set. It is also worth noting that the training loss is calculated after every batch.

According to Fig. 13 and Fig. 14, it showed that the training loss is minimal for the YOLOv5l model. The lesser the training loss, the faster a model's performance can be obtained.



Fig. 13. Train/Box Loss Comparison between Models before Augmentation.

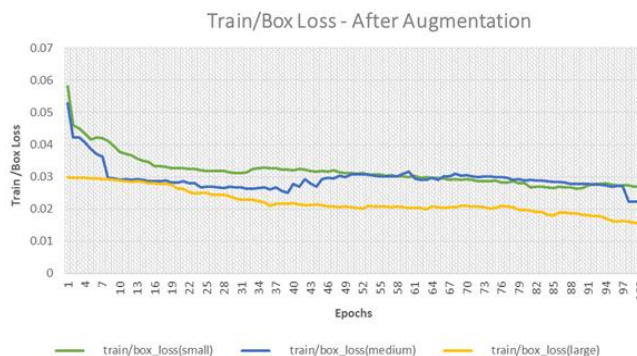


Fig. 14. Train/Box Loss Comparison between Models after Augmentation.

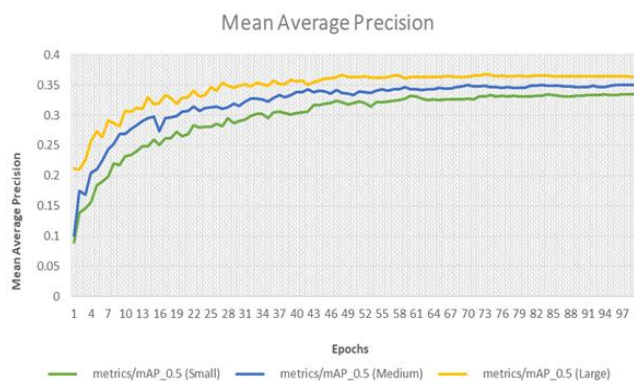


Fig. 11. mAP Comparison between Models before Augmentation.

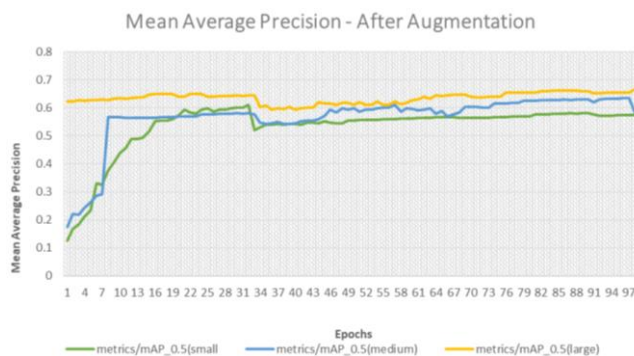


Fig. 12. mAP Comparison between Models after Augmentation.

### C. F1 Score Analysis

By comparing the F1 Score among the three models before and after the augmentation of the dataset as seen in Fig. 15 and Fig. 16, we can see that the large model has a higher score, followed by the medium model and the small model.

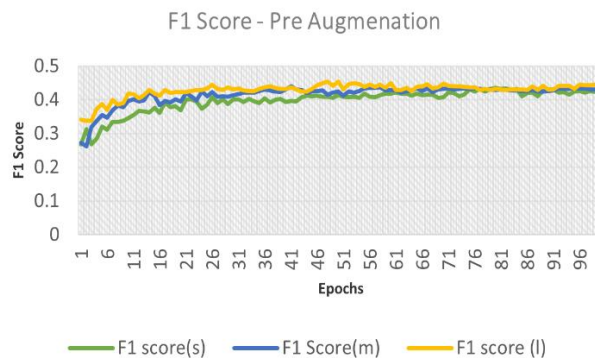


Fig. 15. F1 Score Comparison between Models before Augmentation.

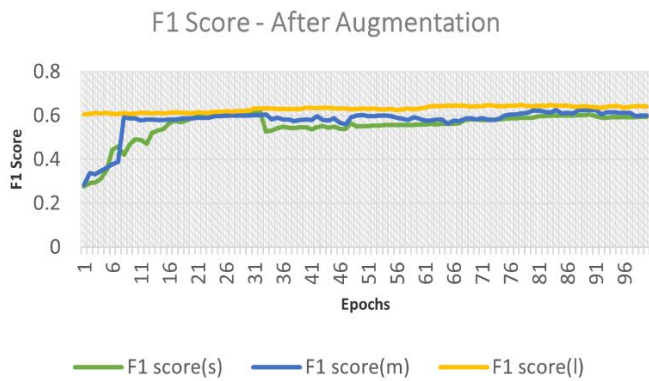


Fig. 16. F1 Score Comparison between Models after Augmentation.

#### D. Augmented Vehicle Dataset Epoch Results

We have augmented our dataset and run for 100 epochs. At the end of the last epoch, the precision was 0.64, recall was 0.63 and mAP was 0.66 for an augmented dataset in YOLOv5 three architecture, Table III.

TABLE III. PERFORMANCE MEASUREMENT AND COMPARISON OF DIFFERENT YOLO MODELS

Model	Precision	Recall	mAP_0.5
YOLOv5l	0.64214	0.63763	0.66877
YOLOv5m	0.59261	0.60812	0.60112
YOLOv5s	0.58038	0.61149	0.57469

The performance is not up to the mark as our dataset was unbalanced and we have a higher amount of car dataset by comparing with other types of vehicles.

#### E. YOLOv5 Overall Performance Analysis

For analyzing the performance of YOLOV5 based on three different architectures, we have measured the performance based on before and after augmentation data. The outcome is presented in Tables IV and V.

Based on the findings, we can conclude that the YOLOv5l model does better out of the three architectures when it comes to vehicle detection for our dataset.

TABLE IV. RESULT COMPARISON OF YOLOV5S, YOLOV5M AND YOLOV5L BEFORE AUGMENTATION

Attributes	YOLOv5s	YOLOv5m	YOLOv5l
mAP_0.5	0.33505	0.35135	0.36361
mAP_0.5:0.95	0.20561	0.23274	0.24556
train/box_loss	0.03026	0.02182	0.01924
train/class_loss	0.01448	0.00737	0.00503
validation/box_loss	0.03443	0.03407	0.03309
validation/class_loss	0.02466	0.02801	0.02853
F1 Score	0.42211	0.43096	0.4438

TABLE V. RESULT COMPARISON OF YOLOV5S, YOLOV5M AND YOLOV5L AFTER AUGMENTATION

Attributes	YOLOv5s	YOLOv5m	YOLOv5l
mAP_0.5	0.57469	0.60112	0.66877
mAP_0.5:0.95	0.41897	0.43868	0.48071
train/box_loss	0.02711	0.0225	0.01574
train/class_loss	0.00977	0.013675	0.00158
validation/box_loss	0.02235	0.022442	0.02112
validation/class_loss	0.00952	0.009342	0.00802
F1 Score	0.59552	0.600264	0.63987

#### VII. CONCLUSION

The application of the different modules of YOLOv5 has significantly improved the detection of vehicular objects in traffic and on the road. The accumulated dataset has provided a vast amount of variety in vehicle classes which led to a richer and more accurate result across all the models. For performance comparison, we utilized different models with different nodes, layers, and speeds. After an extensive data training and processing, period was carried out, it was seen that YOLOv5l had outperformed the rest. Its deeper node complexity and increased number of convolutional layers make it the most efficient in terms of processing, but it was also seen that it takes the most time although by a short margin. The significant performance efficiency of the YOLOv5l model compared to YOLOv5s and YOLOv5m solidifies the real-world application opportunities for the detection of uniquely South Asian vehicles.

#### REFERENCES

- [1] Bangladesh Road and Transport Authority. Vehicles registered in Bangladesh. 2021. <http://www.brta.gov.bd/site/page/74b2a5c3-60cb-4d3c-a699-e2988fed84b2>.
- [2] Paspelava, Computer Vision Object Detection: Challenges faced. 2021.
- [3] C. Snyder, D. Gonzales, M.Do, and T. Ma, "Congestion estimation using traffic cameras," 2019.
- [4] P. Shinde, S. Yadav, S. Rudrake, and P. Kumbhar, "Smart traffic control system using YOLO," Int. Res. J. Eng. Technol, vol. 6, issue. 12, pp. 169-172, 2019.
- [5] Y. Li, and J. Ibanez-Guzman, "Lidar for autonomous driving: The principles, challenges, and trends for automotive lidar and perception systems," IEEE Signal Processing Magazine, vol. 37, issue. 4, pp. 50-61, 2020.
- [6] Ultralytics. YOLOv5: A family of Object Detection Architectures and Models.
- [7] C. Supeshala, Yolo v4 or Yolo V5 or PP-Yolo? Medium. 2020.
- [8] Vehicle detection at intersections by Lidar System. Smart Sensing for Traffic Monitoring, 81-96, 2020.
- [9] J. Zhang, "MASFF: Multiscale Adaptive Spatial Feature Fusion Method for vehicle recognition," Journal of Computers, 33(1), 001-011, 2022.
- [10] K. Kowol, M. Rottmann, S. Bracke, and H. Gottschalk, "YOdar: Uncertainty-based sensor fusion for vehicle detection with camera and radar sensors," Proceedings of the 13th International Conference on Agents and Artificial Intelligence. 2021.
- [11] F. T. Barwell, Vehicle detection. Automation and Control in Transport, 76-83. 1973.

- [12] J. Ciberlin, R. Grbic, N. Teslić, and M. Pilipović, "Object detection and object tracking in front of the vehicle using front view camera," 2019 Zooming Innovation in Consumer Technologies Conference (ZINC), 2019.
- [13] C. Hisham, "Classifier-based approaches for top-down salient object detection," Doctoral thesis, Nanyang Technological University, Singapore, 2017.
- [14] A. Aggarwal, Yolo explained. What is YOLO and How does it work? Medium. 2020.
- [15] L. Tan, T. Huangfu, L. Wu, and W. Chen, "Comparison of yolo v3, faster R-CNN, and SSD for real-time pill identification," Research Square, 2021.
- [16] J. Liu, and D. Zhang, "Research on vehicle object detection algorithm based on improved yolov3 algorithm," Journal of Physics: Conference Series, 2020.
- [17] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.
- [18] G. Chandan, A. Jain, H. Jain, and Mohana, "Real time object detection and tracking using deep learning and openCV," International Conference on Inventive Research in Computing Applications (ICIRCA), 2018.
- [19] Z. Chen et al., "Real time object detection, tracking, and distance and motion estimation based on deep learning: Application to smart mobility," Eighth International Conference on Emerging Security Technologies (EST), 2019.
- [20] J. Kim, S. J. Sung, and S. Park, "Comparison of faster-RCNN, Yolo, and SSD for real-time vehicle type recognition," ICCE-Asia, 2020.
- [21] D. J. Phillips et al., "Real-time prediction of automotive collision risk from monocular video," ArXiv, abs/1902.01293, 2019.
- [22] X. Wang et al., "Vision-based detection and tracking of a mobile ground target using a fixed-wing UAV," International Journal of Advanced Robotic Systems, volume 11, issue 9, page 156, 2014.
- [23] J. Sokalski, T. Breckon, and I. Cowling, "Automatic salient object detection in uav imagery," Proc. of the 25th Int. Unmanned Air Vehicle Systems, pages 1-12, 2010.
- [24] K. Kanistras, G. Martins, M. Rutherford, and K. Valavanis, "Survey of unmanned aerial vehicles (UAVs) for traffic monitoring," Handbook of unmanned aerial vehicles, pages 2643-2666, 2014.
- [25] B. Xiao, and S. Kang, "Development of an image dataset of construction machines for deep learning object detection," Journal of Computing in Civil Engineering, 35(2), 2021.
- [26] B. Zoph et al., "Learning data augmentation strategies for object detection," Computer Vision – ECCV 2020, 566–583. 2020.
- [27] J. Lin, and M. Sun, "A YOLO-based Traffic Counting System," 2018 Conference on Technologies and Applications of Artificial Intelligence (TAAI), 2018.
- [28] J. Tao, H. Wang, X. Zhang, X. Li, and H. Yang, "An object detection system based on YOLO in the traffic scene," The 6th International Conference on Computer Science and Network Technology (ICCSNT), pp. 315-319, 2017.
- [29] A. Corovic, V. Ilic, S. Duric, M. Marijan, and B. Pavkovic, "The real-time detection of traffic participants using YOLO algorithm," 2018 26th Telecommunications Forum (TELFOR), pp.1-4, 2018.
- [30] A. Salarpour, A. Salarpour, M. Fathi, and M. Dezfoulian, "Vehicle tracking using Kalman filter and features," Signal & Image Processing 2, no. 2, 2011.
- [31] H. N. Phan, L.H Pham, D. N. Tran, and S. V. Ha, "Occlusion vehicle detection algorithm in crowded scene for traffic surveillance system," In 2017 International Conference on System Science and Engineering (ICSSE), pp. 215-220. IEEE, 2017.
- [32] L. Ding, Y. Wang, R. Laganière, X. Luo, and S. Fu, "Scale-aware RPN for vehicle detection," In International Symposium on Visual Computing, pp. 487-499. Springer, Cham, 2018.
- [33] U. Nepal, and H. Eslamiat, "Comparing YOLOv3, YOLOv4 and YOLOv5 for autonomous landing spot detection in faulty UAVs," Sensors 22, no. 2, 464, 2022.
- [34] K. Wang, et al., "Panet: Few-shot image semantic segmentation with prototype alignment," In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 9197-9206. 2019.
- [35] A. Syaharuddin, Z. Zainuddin, and Andani, "Multi-pole road sign detection based on faster region-based convolutional neural network (Faster R-CNN)," 2021 International Conference on Artificial Intelligence and Mechatronics Systems (AIMS), 2021.
- [36] L. Jiang, H. Liu, H. Zhu, and G. Zhang, "Improved YOLO v5 with balanced feature pyramid and attention module for traffic sign detection," In MATEC Web of Conferences, vol. 355. EDP Sciences, 2022.
- [37] D. Sinha, and M. El-Sharkawy, "Thin mobilenet: An enhanced mobilenet architecture," In 2019 IEEE 10th annual ubiquitous computing, electronics & mobile communication conference (UEMCON), pp. 0280-0285. IEEE, 2019.

# Multi-method Approach for User Experience of Selfie-taking Mobile Applications

Shahad Aldahri, Reem Alnanih

Computer Science Department, Faculty of Computing and Information Technology  
King Abdulaziz University, Jeddah, Saudi Arabia

**Abstract**—Taking selfies is a popular activity in most social media applications and applying filters/lenses to these selfies has become one of the most demanded features of such applications. This paper aims to design an application for taking selfies to minimize the heavy use of beautifying filters. To understand the current user experience of selfie-taking and filter features, multiple user experience research methods were applied in two steps. In the first step, interviews were conducted with 10 participants to collect data. The key findings of interviews were (i) the need for saving memories as users' primary goal of using the applications, (ii) the need for using slightly beautifying filters as their preferred filter type, (iii) the need for a favorite filters list, and (iv) the need for the opportunity to edit selfies after they are taken. This output of the interviews was used as an input for determining the survey questions in the second step. A total of 340 respondents completed the survey and the findings were consistent with those of the interviews. Further pointing to the rising opportunity for a new selfie-taking application designed to save selfies quickly without sharing and only apply slightly beautifying filters. More studies should focus on increasing engagement and including a saved selfie categorization feature in the design.

**Keywords**—Filters; lenses; multi-method; research methods; selfie taking; user experience research

## I. INTRODUCTION

The Social media platforms have become indispensable Internet-based communication tools for daily life, in part due to the wide range of connectivity they provide. Social media and the Internet have caused the social participation rate to increase rapidly over time; in 2021, the number of Internet users worldwide reached 4.9 billion [1]. Furthermore, social media use is considered one of the most popular online activities [2]. In Saudi Arabia, active social media users represent 79.25% of the population, with an annual growth rate of 8.7% [3]. One of the most popular types of social media platforms are those that provide photo-sharing services—e.g., Instagram and Snapchat—and taking, viewing, and sharing selfies has become a daily habit for many people. Selfies edited with filters are an especially rising trend.

Snapchat is one of the fastest-growing selfie-taking social media applications, with a massive number of users that reached approximately 306 million in 2021 [4]. In 2016, Snapchat introduced a feature called Lenses to offer a range of filters to users, making it the first social media application to utilize augmented reality (AR) technology. Its facial recognition software allows users' mobile cameras to detect

their facial features and overlay a chosen virtual effect on their faces, which alters their facial features to look funny, scary, or more beautiful [5].

The problem with many existing selfie-taking applications that allow users to take selfies with filters is the heavy daily use of filters. People seeing each other from a distance may no longer know what their actual faces look like without filters. Social comparisons are no longer made between actual faces but between one's unfiltered face and the filtered faces of others. This may affect self-evaluation and self-esteem, especially because physical appearance is strongly related to self-esteem [6].

This paper aims to apply user experience (UX) research methods to better understand the current UX of selfie-taking and filter feature in order to minimize the use of heavily beautifying filters to help users accept the reality of their appearance. The UX was investigated through three different views—users' goals of using selfie-taking applications, their requirements for taking selfies, and their behaviors when taking selfies—to answer the following research questions:

- [Goal] Why are people using selfie-taking applications and using the filters feature?
- [Need] What is necessary for taking selfies and using the filters feature?
- [Behavior] How do users take selfies with existing mobile applications?

The strength of this paper lies in its use of several UX methods and techniques. Interviews and surveys were conducted sequentially to collect the requirements data. The findings from the interviews determined what to investigate in the surveys to produce clearer insights. Active listening and affinity diagrams were further used to analyze the collected data.

To the best of the author's knowledge, no existing study has conducted a multi-UX research methods approach to study the UX of selfie-taking and filters in order to minimize the reliance on heavily beautifying filters. The findings of the study should lead designers in this domain to consider the following implications in designing a new selfie-taking application:

- Design a selfie-taking application dedicated to saving memories.



- Allow only slightly and realistically beautifying filters in the application to encourage users to accept their looks without relying on heavily beautifying filters.
- Minimize the time required to take a quick selfie with quick reach, and provide a limited number of customized favorite filters.
- Minimize the time required to edit selfies with quick-adding capabilities for items such as gifs, emojis, text, and automatic date and time.

The rest of the paper is organized as follows: The second section highlights the related work in the field of UX, the third section provides the data collection and analysis, the fourth section presents the results of the study and discussion, and the fifth section concludes the current work.

## II. RELATED WORK

Many researchers have conducted UX studies dedicated to understanding, evaluating, and improving the UX of hardware and software products. One study of hardware aimed to understand the UX of unplanned smartphone use [7] via an ethnographic method that involved video recordings with wearable cameras on the chest to capture spontaneous smartphone use in everyday activities. According to the researchers, unplanned smartphone use leaves no memory of the actions in the past, presents no attributes of what it is, and projects no endings of what it will be in the future. Another study examined the UX of SmartTVs to identify UX factors that influence Smart TV UX factors over different periods [8]. Different UX methods were exploited in different stages, such as surveys, thinking aloud, and daily diaries.

The empirical evidence from each method suggests that the UX factors vary with respect to product temporality. Moreover, UX research has been conducted to design smart, wrist-worn digital jewelry [9]. The authors identified the requirements from the literature and conducted semi-structured interviews with jewelers and potential users. After data analysis and prototyping, they identified the final concept of the digital bracelet and implemented an operable and wearable prototype, which they then used to measure the UX and usability.

The participants' experience was positive. Another researcher used a survey to develop a storyline for an animated video to help generate empathy from the viewer for people with social anxiety disorder [10]. The viewers of the video were able to empathize with the main character in the animation even though the story did not have a defined ending.

Many UX studies have contributed to the UX of mobile applications and websites. For example, the gendered perceptions of and initial UX of Pinterest were studied to improve the binary disparity in its use between men and women [11]. The results of the surveys conducted in the study revealed significant differences in perceptions between users and non-users and between men and women. Another study evaluated the UX of Snapchat's onboarding process,

using a new multi-method approach that could be used to evaluate UX for any mobile application [12]. The new methodology consisted of three methods: a usability test interview, an affinity diagram, and a model of UX attributes. The UX attributes were applied to Hassenzahl's model to define which were most important and identify significant development points for the application management team. One researcher evaluated Facebook's UX and investigated the differences in experiences between frequent users and new users using electroencephalography (EEG) [13]. Their findings present a significant statistical difference between new and frequent Facebook users.

Table I summarizes the published date, type of studied product, UX method used, and domain of these studies.

The achievements of the above studies highlight that UX research can contribute to a theoretical background for the development and design of products and analysis of user behavior for various products and domains. Furthermore, different UX methods can be used to collect and analyze data in UX research to deliver reliable findings, and mixed methods can be used to gain more knowledge about UX [8][9]. The Smart TV study used different methods in different stages (i.e., before usage, during usage, after usage). Similarly, the digital bracelet study used one method before usage and another method during usage.

However, none of the above studies used 1) mixed methods in a single stage or 2) sequential methods where findings from the first method determined what to measure in the second method. Moreover, the only research that studied the UX of social media with selfie-taking focused only on the Snapchat UX in the onboarding process, and it failed to discuss Snapchat filters [12].

TABLE I. RELATED UX STUDIES

Study	Date	Studied Products	UX Methods	Domain	Hardware/Software
[7]	2016	Mobile phone	Video Recordings	Social	Hardware
[8]	2016	Smart TV	Survey, Think Aloud, Daily Diary		
[9]	2016	Wearable	Interview, Live User Observation		
[10]	2019	Video	Survey	Mental Health	Software
[11]	2015	Website	Survey	Social Media	
[12]	2020	Mobile Application	Interview		
[13]	2020	Mobile Application	EEG		

## III. DATA COLLECTION AND ANALYSIS

Multi-sequential methods of qualitative and quantitative were conducted to investigate the topic in more depth. The 2-step process of data collection and analysis is shown in Fig. 1.



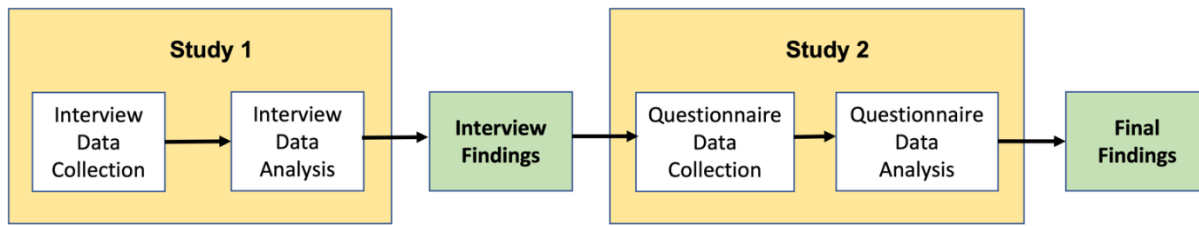


Fig. 1. Two-step Data Collection and Analysis.

The first step of the methodology included interview data collection and analysis. Interviews were conducted first to develop a more thorough understanding with a small number of users. The outcome of the interviews and data analysis informed the topics explored in the survey, which were narrowed in scope to just a few key findings to allow for a deeper investigation. The second step was survey data collection and analysis, which sought to determine whether the interview key findings applied to a broader subset of Snapchat users.

#### A. Step 1: Interview Data Collection and Analysis

The interviews were designed with 9 opening questions to elicit initial responses to a specific topic, providing a useful springboard for follow-up questions. The questions explored the participants' demographic information, previous and current usage of selfie-taking apps, goals and motivations of using these apps, pain points and needs within the apps, and contexts for taking selfies with and without filters. Recruiting of interviewees ensured they met the following criteria: selfie-taking applications used at least once a week, filters used at least once a week, and at least 1 month of previous use of selfie-taking applications.

Ten participants were recruited for 30-minute interviews, which were conducted in the summer of 2021. Nine participants were interviewed by phone due to circumstances related to the COVID-19 pandemic and one was interviewed in person. Six participants were women and four were men. Their ages ranged from 19 to 32 years, and each had used some combination of Snapchat, Instagram, and TikTok for taking selfies.

The data collected from the interviews were analyzed using two UX methods: active listening and an affinity diagram.

1) *Active listening*: The active listening technique condensed the large amount of data gained from the interviews into smaller units that were easier to understand [14]. This technique involved listening to the important things said by the interviewees, then paraphrasing each response into a single quote that captured its essence. The digital tool Miro was used for this task [15]. The following sentences fragments are examples of the small-unit responses resulting from active listening of the interviews with Participants 1, 2, and 4:

- Participant 1: "I would like to reach my favorite filters quickly".

- Participant 2: "I use Snapchat 90% for saving memories".
- Participant 4: "I like using filters that beautify me but allow me to look like myself".

2) *Affinity diagram*: The affinity diagram technique was used to organize and analyze the large amount of data gathered from the interviews. Its purpose is to arrange many pieces of data into manageable clusters or groups that identify patterns in the data [16][17]. To construct the affinity diagram, the responses were categorized into three main groups—user behavior, user needs, and filters that users like or dislike—and further subcategorized into a total of 9 specific themes, as shown in Fig. 2.

The key findings of the affinity diagram are listed below. The finding names (e.g., F.A, F.B) correspond to Finding A, Finding B, etc., and were used to simplify the process of mapping the key findings from Step 1 to the survey questions in Step 2.

These findings are accompanied by supporting evidence, such as screenshots, quotes from the participants, and the number of participants that agreed on the finding.

Key findings:

F.A: Users take selfies primarily to save memories.

- "I use Snapchat 90% for saving memories" Participant 2.
- "Saving memories is my first goal, then viewing others' selfies." Participant 3.
- "I usually take selfies with others, mostly for saving not for sharing." Participant 7.

F.B: Users prefer slightly and realistically beautifying filters over beautifying filters that are unrealistic and over the top.

- Most participants (7 of 10) preferred to still look like themselves after using beautifying filters. They use them to look healthier, with smoother skin and a few realistic changes to their facial features.
- "I hate the beautifying filters that change me to look like everything but me, like I am another person" Participant 2.
- "I like filters that only give me a healthy look" Participant 5.

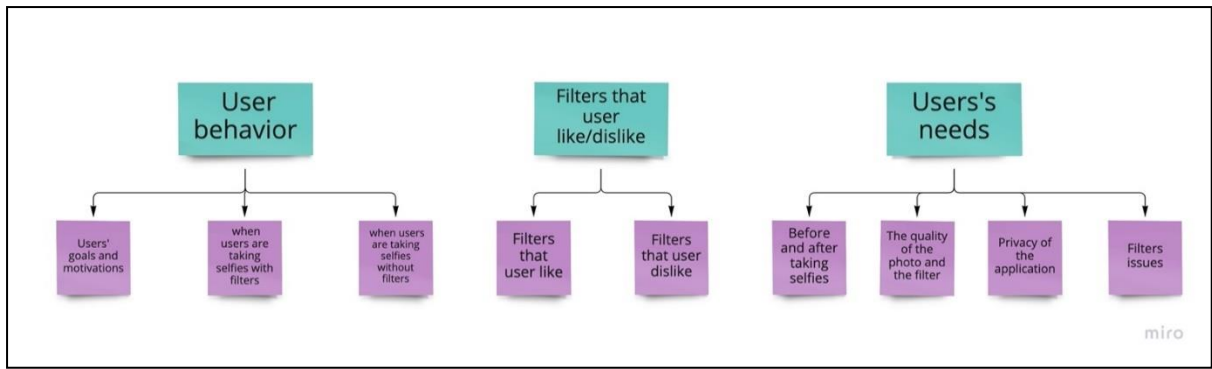


Fig. 2. Levels of Themes of Affinity Diagram.

F.C: Users want to reach their favorite filters quickly before taking selfies.

- “I would like to choose from my list of filters that I can reach quickly” Participant 6.
- “It would be helpful if only the favorite filters appear with one tap before taking the selfie” Participant 10.
- “I like that in Snapchat I can put stars on my favorite filters, but I have to go to explore filters to find my favorite filters” Participant 4, see number 1 and 2 in Fig. 3.

F.D: Users like adding items to their selfies, such as gifs, emojis, text, and the date and time.

- “I enjoy using gifs and emojis on selfies” Participant 1.
- “I like that Snapchat saves the date and time automatically” Participant 3.
- “I like to add music, stickers, text, time, and date to the selfies” Participant 6.

**B. Step 2: Survey Data Collection and Analysis**

The survey was divided into three sections: demographic questions, screening questions, and requirements questions. The demographic questions gathered demographic information that may affect users’ selfie-taking experience. The screening questions ensured that the respondents met the following qualifying criteria: selfie-taking applications used at least once a month, filters used at least once a month, and the device used for taking selfies is the mobile. Finally, five requirements questions were posed to investigate the key findings of the interviews. Table II presents the requirements questions, the corresponding key findings, and the rationale for the selection of these questions.

The survey data were collected in November 2021 in Saudi Arabia. The final sample consists of 460 participants. The quantitative data were analyzed using statistical procedures with SPSS. Descriptive statistics and charts were used.

1) *Demographic questions analysis:* Most respondents were in the 21–30 age group and the 10–20 age group had the fewest number of respondents. Furthermore, 35.2% were

men, 64.8% were women, and 81.5% held bachelor’s or graduate degrees.

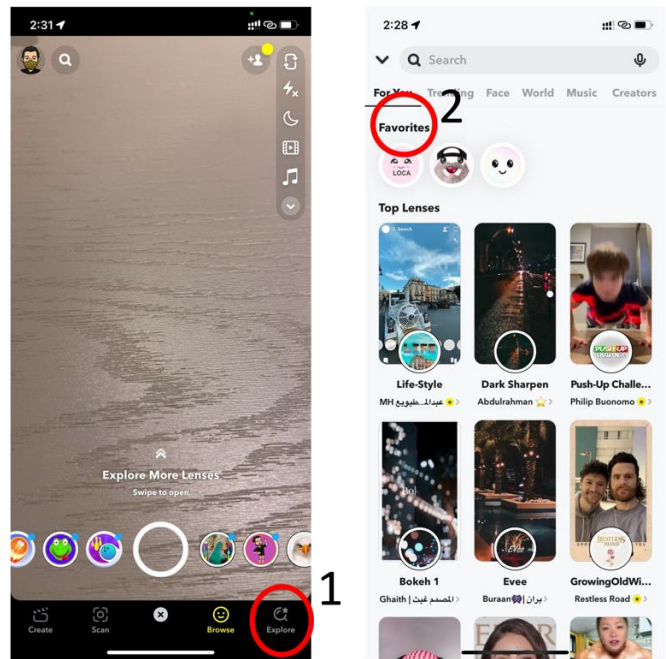


Fig. 3. Reaching Favorite Filters in Snapchat.

TABLE II. REQUIREMENTS QUESTIONS

Requirements Questions	Key Finding(s)	Rationale
Which of the following is your goal of taking selfies?:	F.A	Understand the user’s goal
Which type of filter do you prefer to use?	F.B	Understand the user’s preference of filter types
Do you have any favorite filters that you usually use?	F.C	Identify if the user has a favorite filters list
How many favorite filters do you usually use?	F.C	Identify the usual number of favorite filters for users
Based on your experience with the selfie-taking app(s) you use, how essential were the following features?:	F.C , F.D	Determine what is necessary for users before and after taking the selfies

2) *Screening questions analysis:* The screening questions filtered out 120 respondents, leaving 340 to complete the survey.

Table III shows the percentages of male and female respondents who were frequent users of filters (i.e., used filters daily or weekly) and those who were infrequent users of filters (i.e., used filters monthly or less).

Most respondents (53.55%) used filters daily or weekly. The 21–30 age group had the highest daily usage of filters in selfie-taking applications, and women had higher daily filter usage than men.

TABLE III. FREQUENCY OF USING FILTERS BY GENDER

Gender	Frequent use of filters	Infrequent use of filters
Male	42.86%	57.14%
Female	59.12%	40.83%
All	53.55%	46.45%

3) *Requirements questions analysis:* The selfie-taking application used most often was Snapchat with 95.6% of responses. The results of the multiple-choice questions regarding the users' goals of using selfie-taking applications revealed saving personal memories as the most common goal (71.1%, n=241), followed by sharing with family and friends (64.9%, n=220). For most age groups and both men and women, saving personal memories was the most chosen goal; the exception is respondents over 40 years old, for whom sharing with family and friends was most chosen. Seeing one's face with a filter was the third most chosen goal, mostly by female respondents (76.62%, n=118) and respondents who used filters daily. Sharing with public users was the least chosen goal (9.4%, n=32); male respondents chose this goal more than female respondents. Fig. 4 shows the distribution of goals for each age group, and Fig. 5 shows these goals for men and women.

For the questions on filter type preferences, most of the respondents chose the slightly beautifying filter type (83%, n=284). The advanced beautifying filter type was chosen mostly by respondents who used filters daily.

The majority of respondents (58.1%) had favorite filters that they usually used, whereas the other 41.9% did not. The most common numbers of favorite filters were 3 (28.4%, n=56), 2 (24.4%, n=48) and 4 (19.8%, n=39).

Cronbach's  $\alpha$  (0.71) was calculated to test the reliability of the responses to the five questions on the essentiality of different features in selfie-taking applications. Most respondents (55.16%) considered accessing favorite filters quickly to be a must-have feature and 35.40% considered it a nice to have feature, as shown in Fig. 6.

Adding text and date features were mostly considered nice to have features. Adding the time, emojis, and gifs were mostly considered unnecessary or nice to have features.

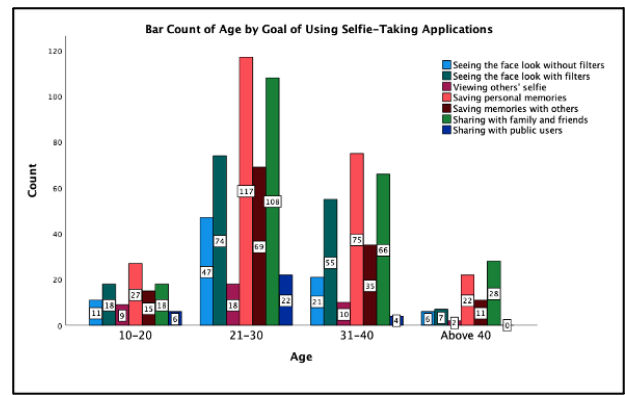


Fig. 4. Goal of using Selfie-taking Applications for Each Age Group.

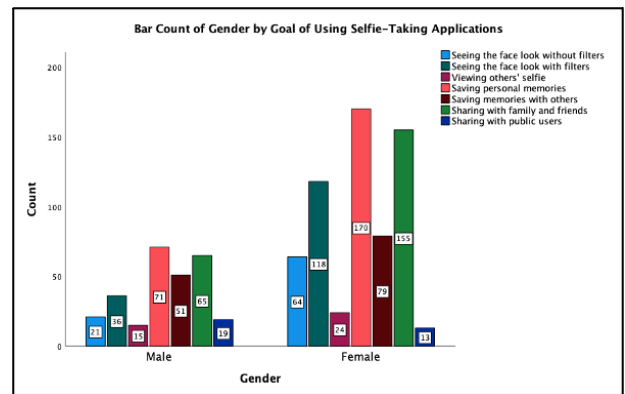


Fig. 5. Goals of using Selfie-taking Applications for Men and Women.

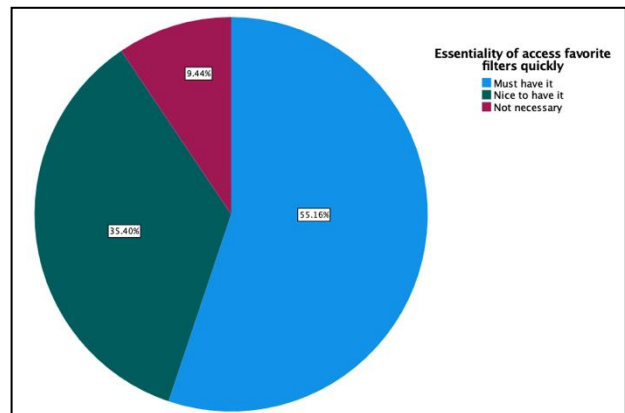


Fig. 6. Percentage of Essentiality of Accessing Favorite Filters Quickly.

#### IV. RESULT AND DISCUSSION

By focusing on areas that were rich with respondents, the final key findings were identified and supported by evidence.

Final key findings are:

- Users prefer slightly and realistically beautifying filters over unrealistically and excessively beautifying filters.
- 83.8% of participants preferred slightly beautifying filters and only 15.6% preferred advanced beautifying filters.

- Users enjoyed adding text, the date and/or time, and emojis to their selfies.
- 75.2% of participants considered adding text as either a must-have or nice to have feature.
- 67.3% of participants considered adding the date as either a must-have or nice to have feature.
- 59% of participants considered adding the time as either a must-have or nice to have feature.
- 58.4% of participants considered adding emojis and gifs as either must-have or nice-to-have features.
- Users had favorite filters and wanted to reach them quickly before taking selfies.
- 58.1% of participants had favorite filters they usually used.
- 55.16% of participants considered quick access to their favorite filters as a must-have feature.
- Users take selfies primarily to save memories.
- 71.1% of participants chose saving memories as their goal of taking selfies.
- Saving memories was the most chosen goal for both men and women and all age groups, except >40 years.

The above result shows that the key findings of the survey align with those of the interviews. From this, an opportunity was identified for a new selfie-taking application that differs from the existing selfie-taking applications. The main points considered for the new application were:

- The ability to take a selfie just for saving, not for sharing.
- The ability to reach one's favorite filters with one click before taking selfies.
- A collection of slightly and realistically beautifying filters.
- The ability to add the time and date, gifs, emojis, and text on taken selfies.

Each research method has both positive and negative attributes. Surveys and interviews are two of the most used methods in UX studies. However, using a survey alone leads to a disconnection between the researcher and respondents to observe emotional concerns that may be evident in respondents' responses. Also, no possibility of correcting the misunderstanding which may be due to the wording or ambiguity in the survey. In other hand, using interviews alone has several cons such as sex, race, and class affiliation of the interviewee may play a role in the bias of the interviewer. Also, recruiting a large sample in the interview method will require higher costs and time. The limitation of a single-method approach was overcome by employing multiple research methods and the double diamond framework to understand the full UX. The sequential usage of these research methods allowed the researchers for a more in-depth investigation of the research topic. Rather than using the

survey to investigate every point or issue from the interviews, the scope was narrowed to the key findings of the interview that merited further investigation. The survey therefore focused on users' goals of using selfie-taking applications, their preferences for filter types, the need for a favorite filters list, and the essential features for taking selfies. This multi-sequential approach proved able to provide consistent findings.

The existing studies had a limitation in evaluating the UX of selfie-taking applications, especially the evaluation of filters feature. The contribution of this paper is using the benefits of multi-sequential research methods and the strength of the double diamond framework to evaluate the UX of selfie-taking and filters.

People's daily reliance on heavily beautifying filters motivated this research. However, the results of the research methods illuminated two new problems in selfie-taking applications. The first is that most selfie-taking applications focus on sharing selfies with other users because, like Snapchat, they are considered social media platforms. However, most of the participants of both the interviews and surveys chose saving memories as their goal in using these apps, which answers question 1 of this study. The second problem is that the excessive numbers of filters available mean that users must search through them to choose one for each selfie; this extends the time required to take a quick selfie. Many respondents expressed the need for a quick reach time to the favorite filters list, which answers question 2 of this study. Because people live busy lives, they sometimes wish to stop only for a second to take a quick and almost neutral selfie with slightly enhancing beauty—just to save the memory for themselves, not to share, which answers question 3 of this study. A new selfie-taking application could solve these problems by focusing solely on taking selfies instead of sharing them and by offering slightly beautifying filters and the option to quickly save the memories.

## V. CONCLUSION

Recently, applications that support people's interest in taking selfies by providing various features, such as filters, are in high demand. However, the consequences of daily reliance on beautifying filters have increased. Thus, to understand the current UX of selfie-taking applications, this empirical study investigated the goals, needs, and behaviors of their users in order to find opportunities for improvement and solve the rising reliance on heavily beautifying filters. The multi-sequential methods approach was conducted to collect and analyze data on two steps. The result of these analyses revealed an opportunity for a new selfie-taking application that focuses on taking selfies to save memories rather than sharing them, providing only slightly beautifying filters rather than heavily beautifying filters, and improving the speed and ease of the selfie-taking process to make it compatible with people's busy lives.

The multi-sequential approach of qualitative and quantitative methods combined is expected to be easily generalizable to similar applications and products. Also, this research can serve as guidelines for building new social applications with selfie-taking and filters features. Future

research can build on this study to determine new features for categorizing and handling the saved memories and increasing the level of engagement.

#### REFERENCES

- [1] “Number of internet users worldwide 2021 | Statista.” <https://www.statista.com/statistics/273018/number-of-internet-users-worldwide/#statisticContainer> (accessed Jan. 22, 2022).
- [2] “Top mobile-first activities 2019 | Statista | Statista.” <https://www.statista.com/statistics/783357/leading-mobile-first-activities/> (accessed Nov. 11, 2020).
- [3] “SAUDI ARABIA SOCIAL MEDIA STATISTICS 2021 - Official GMI Blog.” <https://www.globalmediainsight.com/blog/saudi-arabia-social-media-statistics/> (accessed Jan. 22, 2022).
- [4] “Snapchat: daily active users worldwide | Statista.” <https://www.statista.com/statistics/545967/snapchat-app-dau/> (accessed Jan. 22, 2022).
- [5] C. Allsteadt, “An Exploration into the Effect of Advancing Technology on UX of Social Media Applications I . Introduction II . Literature Review,” *Elon J. Undergrad. Res. Commun.* Vol. 8, No. 2, pp. 121–130, 2017.
- [6] “Physical appearance is positively related to self-esteem level.” <https://www.ukessays.com/essays/psychology/appearance-is-positively-related-to-self-esteem-level-psychology-essay.php> (accessed Jan. 08, 2021).
- [7] J. Cousins, “Timelessness,” *Self Soc.*, vol. 15, no. 5, pp. 216–216, 2016, doi: 10.1080/03060497.1987.11084862.
- [8] J. Jang, D. Zhao, W. Hong, Y. Park, and M. Y. Yi, “Uncovering the Underlying Factors of Smart TV UX over Time,” pp. 3–12, 2016, doi: 10.1145/2932206.2932207.
- [9] J. Fortmann, E. Root, S. Boll, and W. Heuten, “Tangible Apps Bracelet: Designing modular wrist-worn Digital Jewellery for multiple purposes,” in *DIS 2016 - Proceedings of the 2016 ACM Conference on Designing Interactive Systems: Fuse*, Jun. 2016, pp. 841–852, doi: 10.1145/2901790.2901838.
- [10] N. Andalibi, “Development of a 2D Animated Video Using UX Research Methods to Generate Empathy for People with Social Anxiety Disorder,” p. 1, 2019.
- [11] H. Miller, S. Chang, and L. Terveen, “‘i love this site!’ vs. ‘it’s a little girly’: Perceptions of and Initial User Experience with Pinterest,” *CSCW 2015 - Proc. 2015 ACM Int. Conf. Comput. Coop. Work Soc. Comput.*, pp. 1728–1740, 2015, doi: 10.1145/2675133.2675269.
- [12] K. Kapusy and E. Lógó, “User Experience Evaluation Methodology in the Onboarding Process: Snapchat Case Study,” *Ergon. Des.*, pp. 1–7, 2020, doi: 10.1177/1064804620962270.
- [13] R. S. Mangion, L. Garg, G. Garg, and O. Falzon, “Emotional Testing on Facebook’s User Experience,” *IEEE Access*, vol. 8, pp. 58250–58259, 2020, doi: 10.1109/ACCESS.2020.2981418.
- [14] C. R. Rogers and R. E. Farson, “Active listening.” Chicago, IL, 1957.
- [15] “The Visual Collaboration Platform for Every Team | Miro.” <https://miro.com/index/> (accessed Feb. 14, 2022).
- [16] C.-K. Kwong and H. Bai, “Determining the importance weights for the customer requirements in QFD using a fuzzy AHP with an extent analysis approach,” *Iie Trans.*, vol. 35, no. 7, pp. 619–626, 2003.
- [17] B. Martin, B. Hanington, and B. M. Hanington, “Universal methods of design: 100 ways to research complex problems,” *Dev. Innov. Ideas, Des. Eff. Solut.*, pp. 12–13, 2012.

# Predicting Academic Performance using a Multiclassification Model: Case Study

Alfredo Daza Vergaray<sup>1</sup>, Carlos Guerra<sup>2</sup>, Noemi Cervera<sup>3</sup>, Erwin Burgos<sup>4</sup>

Professor, School of Systems and Computer Engineering, Universidad Nacional del Santa, Ancash, Perú<sup>1,2</sup>

School of Systems and Computer Engineering, Universidad Nacional del Santa, Ancash, Perú<sup>3,4</sup>

**Abstract**—Now-a-days predicting the academic performance of students is increasingly possible, thanks to the constant use of computer systems that store a large amount of student information. Machine learning uses this information to achieve big goals, such as predicting whether or not a student will pass a course. The main purpose of the work was to make a multiclassifier model that exceeds the results obtained from the machine learning models used independently. For the development of our proposed predictive model, the methodology was used, which consists of several phases. For the first step, 557 records with 25 characteristics related to academic performance were selected, then the preprocessing was applied to said data set, eliminating the attributes with the lowest correlation and those records with inconsistencies, leaving 500 records and 9 attributes. For the transformation, it was necessary to convert categorical to numerical data of four attributes, being the following: SEX, ESTATUS\_lab\_padre, ESTATUS\_lab\_madre and CONDITION. Having the data set clean, we proceeded to balance the data, where 1,167 data were generated, using the 2/3 for training and the remaining 1/3 for validation, then the following techniques were applied: Extra Tree, Random Forest, Decision Tree, Ada Boost and XGBoost, each obtained an accuracy of 57.41%, 61.96%, 91.44%, 59.65% and 83.3% respectively. Then the proposed model was applied, combining the five algorithms mentioned above, which reached an accuracy of 92.86%, concluding that the proposed model provides better accuracy than when the models are used independently meaning that it was the one that obtained the best result.

**Keywords**—Learning machine; prediction; academic performance; hybrid model; classification techniques; multiclassification; python

## I. INTRODUCTION

The performance of a student, over the years, has always been of great importance to the institutions that provide teaching, which is why much research is done on academic achievement.

On the latter, [1] he states that "it is of great importance to support the development of students and improve the quality of higher education, which ultimately improves the reputation of institutions" (p. 21). Therefore, education plays a very important role in the progress of any society, where learning outcomes are seen as an indicator related to better health, social and more effective careers and a factor of improvement of families and communities [2].

According to [3] they indicate that about 25% of every 100 students at the higher level abandon their training in the first semester. Most of them start with failed subjects and low

averages, in the third semester there is a dropout rate of 36%, a figure that increases semester by semester, until reaching 46%, which makes academic performance very transcendental and important. These results show that today's young people have the minimum of the skills needed to perform capable in contemporary societies; they have serious deficiencies to start their professional studies and of course they will have serious problems to successfully insert themselves both into the labor market and into the social, scientific, political and business groups that run the country.

Likewise, universities undertake to update their study plans and programs to adapt them to the needs of today's society; Unfortunately, while these efforts are important, modifying or changing the curriculum does not eliminate learning problems, but also presents new challenges. Similarly, according to [4] to consider a university as one of high quality it is necessary that it has an excellent record of academic achievement.

As an idea of solution to achieve this goal, is that the use of new technologies is becoming more and more frequent, as is the case of data mining.

According to [5] "Data mining is a process of automatically extracting useful information from large data set repositories, etc." In addition, the usefulness of this technology is that it can be used to train learning models, which from historical data can discover useful learning information and based on this, make a prediction [6].

Currently this technology is applied in various fields such as industry, banking, among others. Applied to the field of education, it is called Educational Data Mining (EDM), which, according to [7] "is an emerging area of research composed of a large set of psychological and computational approaches to provide a roadmap of how students learn." On the other hand, at present the existence and constant use of automated learning tools allow, according to [7] to store "a variety of data related to students and valuable characteristics that affect the performance of students and that can be used in the construction of the prediction model".

In Peru, the problem of low performance is frequent at different levels of education. The performance of the students of the different engineering faculties of the Universidad Nacional del Santa is medium low of the vast majority according to the report made by the [8], this affects both the university, since "its success depends on the success of its students" according to [9], and the students themselves, since



they opt for desertion, career change, or limit them when it comes to finding work in the future, where in their studio they developed a predictive model using the J48 technique, which obtained an accuracy of 60.9%.

Currently there are many learning models that are used to extract knowledge from a data set, some of these are those based on Naive Bayes (NB), Vector Support Machine (SVM), Decision Trees (DT), the Closest Neighbor (KNN), among others; but as stated [10], it is difficult to find an efficient classification model that can be used for various situations or problems. That is why the idea of combining several classifiers (multiclassifiers) was born.

The multiclassifiers according to [18] "belong to a recent area of data mining that has allowed to improve, in general, the accuracy of predictions through the combination of individual classifiers" and some of these multiclassifiers are Streaming Ensemble Algorithm (SEA), Coverage Based Ensemble Algorithm (CBEA), multiclassification based on CIDIM (MultiCIDIM-DS) and MultiCIDIM-DS-CFC.

Based on the problem that arises in the university and in search of improving the various solutions proposed by several studies, it was proposed to create a hybrid model that is capable of predicting academic performance so that students and teachers can opt for preventive measures to avoid that grades are deficient in the future.

This article aims to create a predictive model making use of multiclassification through the Stacking technique using new algorithms such as: Extra Tree, Random Forest, Decision Tree, Ada Boost and XGBoost to achieve better accuracy, taking into account that the studies that have been done so far, make use of a single prediction technique, thus generating a good precision, but that could be better if several techniques were applied together.

The rest of the work is structured in five sections. In Section II, a review of the literature of related works is presented. Section III contains the method, which outline the data mining process implemented in this study, which includes a representation of the collected dataset, an exploration and visualization of the data, and finally the implementation of the data mining tasks and the final results. Section IV shows the findings and discussion obtained after the creation and testing of the predictive model. Finally, Section V contains the conclusions reached after the development of the model.

## II. RELATED WORK

This section compiles various research conducted in recent years on the application of data mining in the education sector.

There are many studies or works carried out related to education, whose main theme is the academic performance or dropout of students such as the study carried out by [8] whose objective was to compare various data mining techniques when using them to predict the performance of students. The data they used was from the Kaggle repository and this comparison included the techniques: Decision Tree (C5.0), Naïve Bayes, Random Forest, Support Vector Machine, K-Nearest Neighbor and Deep Neural Network, this being the one that obtained a greater precision, reaching 84% accuracy.

Another work is also carried out by [11], aimed to perform a comparison and study of hybrid classification model and machine learning algorithms based on decision tree, clustering, artificial neural network, Naïve Bayes, etc., using the open source data mining tool Weka for a practical experiment on a student dataset, having as results that the hybrid method achieved the highest accuracy of 92.59% than individual classifiers, that is, J48, NB, IBK and ANN achieved an accuracy of 85.18, 81.48, 88.88 and 88.88%.

In [12], the authors developed a regression model to predict the score that a student would have, used the ALGORITHM KNN, Decision Tree, SVM, Random Forest and Multiple Linear Regression. After comparing the results of each algorithm, it was the Multiple Linear Regression Model that obtained the greatest accuracy. In [13], They conducted a study of student dropout to determine what were the causative factors and which classification algorithm is the most used to predict this problem. After reviewing several studies, they concluded that decision tree classifiers were the most commonly used, as they obtained good predictions.

A semi-supervised learning approach is the one they used [14] to rank the performance of first-year college students. The categories to classify were low, medium and high and the classifier was Naive Bayes, who obtained an accuracy of 96% and specificity of 100%.

In [15], they build a model that predicts the outcomes students will achieve in the semester. They used 13 learning algorithms, belonging to 5 categories, for the Bayes category, they used Naive Bayes, for the Function category they used SVM and Perceptron Multilayer, for the Lazy category, the IBK technique, for the Rules category they used Decision Table, JRip, OneR, Part and ZeroR and finally for the Trees category, they used the J48 techniques, Random Forest, Random Tree, and Simple CART. The data correspond to 50 students and they developed their model on the Weka platform, after the results of having applied each of these techniques, the one that had the best results was the J48 technique, which reached 88% accuracy in the prediction

In the same way the study carried out by [16], whose objective is to develop a prediction model based on Bayes, specifically Naive Bayes and Bayes Network. The data was collected through a questionnaire of 62 questions related to health, social activity, relationships and academic performance. They used the Weka tool in which they obtained as a result the algorithm Naive Bayes is better, since it obtained 70.6%, while Bayes Network obtained 64.3% accuracy.

In [17], the author in his research aimed to develop an algorithm with incremental learning to mine data flows that is capable of manipulating gradual, abrupt or recurrent concept changes, obtaining as results that the FAE algorithm achieved promising results in the tests, compared to well-known algorithms implemented also in the MOA work environment, taking into account the parameters: Accuracy (82.4%), execution time, behavior in the transition period from one concept to another and recovery time after a change of concept.

In the present study, five classification algorithms are used as the basis for the creation of a multiclassifier model, through

the Stacking technique, these algorithms being: Extra trees, Random Forest, Decision Tree, Ada Boost, XGBoost.

#### A. Extra Trees

It is the short name of Extremely Randomized Trees, which means Extremely Randomized Trees. This technique consists of a large number of individual decision trees. It is characterized because it uses the entire set of training data, to grow each decision tree [18].

The Extra Trees algorithm creates many decision trees at random, with the intention of finding a final answer, from the combination of the results of each tree. The difference with the Random Forest algorithm, which has the same procedure, is that the number of random processes used in Extra Trees is much higher.

#### B. Random Forest

A random forest consists of many decision trees. Each tree in the forest is a binary tree and its generation follows the principle of top-down recursive division [19]. For each tree, the root node contains all the training data and this is divided into two nodes, the left and right, according to certain rules and these in turn train with different samples of data. The division continues to occur based on certain rules until the fork stop is met.

#### C. Decision Tree

The decision tree is a tree-like structure that represents a series of decisions and the resulting decision takes the form of rules for classifying a given data set [20]; these are supervised algorithms that can be used for both classification and regression. The objective of this algorithm is to predict by learning decision rules. After the construction of a decision tree, these classify an instance from the root node of the tree then it is directed to a leaf of the tree that would be the intermediate node, depending on the value it takes and this is done successively until it reaches the last leaf of the tree that would be the terminal node.

#### D. Ada Boost

The Ada Boost algorithm stands for Adaptive Boosting and is one of the most popular techniques of the Boosting method. This algorithm is iterative and its operation consists of training different classifiers considered weak for the same set of training data, then combining them to form a stronger classifier [21].

Ada Boost classifiers represent a robust class of classifiers that aim to increase or improve the accuracy of an already built classifier [14].

#### E. XGBoost

XGBoost is an advanced software based on Gradient Tree Boosting that can efficiently handle large-scale machine learning tasks [22]. The XGBoost algorithm stands for Extreme Gradient Boosting and is a supervised algorithm based on the Boosting method. To achieve the strongest classifier, an optimization algorithm is used, Gradient Descent and each model generated is compared with the previous one and if a new model obtains better accuracies, this is taken as a basis for relicensing modifications. But, if in case, its accuracies are low, it returns to the previous model, making modifications based on this. The process is repeated until the differences between two consecutive models are negligible, which means that the maximum number of iterations was reached.

#### F. Stacking

Set learning algorithms are metaalgorithms that combine different machine learning algorithms into a single predictive model to reduce bias (boosting), variance (bagging) or improve the accuracy of predictions (stacking) [23]. Based on the study presented in this section a stacking method was developed (see Fig. 1). This technique involves using predictions from previous-level machine learning models as input variables for next-level models [5].

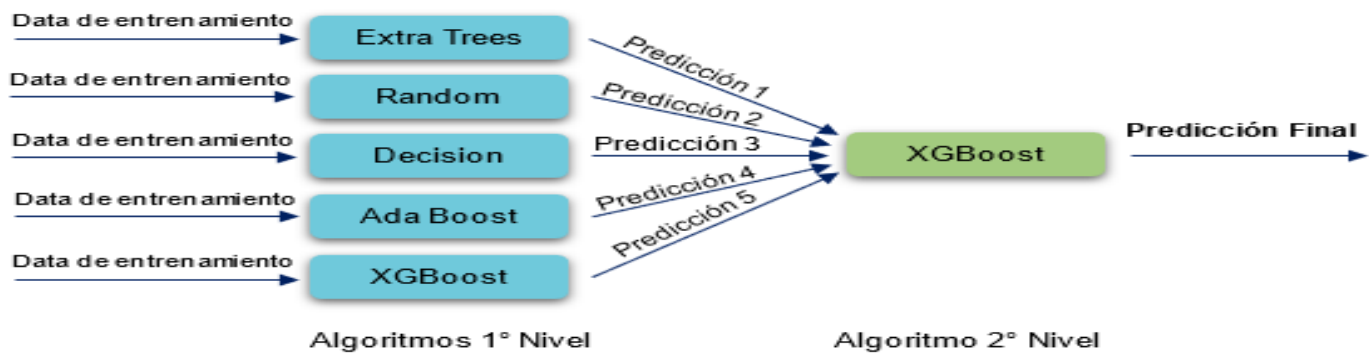


Fig. 1. Graphical Representation of the Stacking Method used in this Work.

### III. METHOD

For the development of the present study, the steps of the stacking method shown in Fig. 2 have been followed, it should be noted that XGBoost is an advanced software based on Gradient Tree Boosting that can efficiently handle large-scale machine learning tasks [14].

1) *Data integration*: The data of systems and computer engineering and agroindustrial engineering were integrated.

2) *Selection*: This work focuses on students of systems engineering, energy and agro-industrial of the Universidad Nacional del Santa. Data were collected using an online questionnaire, which included questions related to some characteristics about academic performance. The questionnaire contained a total of 25 characteristics and were answered by 557 students, of whom 135 were female and 422 were male. The characteristics considered in the questionnaire are detailed in Table I.

3) *Pre-processing*: This step is important to prepare the data before it is used in testing. In our case, the data required pre-processing, as there was empty data, inaccurate data and irregular or inconsistent data. Some of the tasks included in this step are: data cleansing, transformation data, reduction data, and integration data [19]. Another detail of the collected data set is that they are mostly categorical, and for this data to be used in the selected tool, Python, it must necessarily be numerical data.

a) *Removing attributes*: Initially, the characteristics SCHOOL, Cod\_student, CI\_ante, prom\_trans, NATIONALITY, FECH\_nac, RACE, TYPE\_viv, PLACE\_res and other attributes that do not necessarily have a great correlation with the student's performance were eliminated, as shown in Table II.

b) *Data cleansing*: Data cleansing required deleting records that contained empty or inconsistent data. In the first instance, there were 88 records that had at least one empty value, after their elimination, there was a record that contained an invalid data, finally leaving 500 records to be used in the model to be proposed.

c) *Creating the output class*: The focus of this report is classification, and taking into account that the collected data set had an attribute, AVERAGE\_ACU, which contained the academic averages of the students surveyed, the creation of 3 categories was considered so that the model can classify a certain student in one of those categories. These three categories were considered based on the following:

- Bad, whose rating is less than 10.5.
- Regular, whose rating is greater than or equal to 10.5, but less than 14.
- Regular, whose rating is greater than or equal to 10.5, but less than 14.
- Well, whose rating is greater than or equal to 14, but less than or equal to 20, this value being the maximum in Peru's rating system (vigesimal system).

Therefore, the final output class considered for the classification model is, CONDITION, the following attributes as shown in Table III.

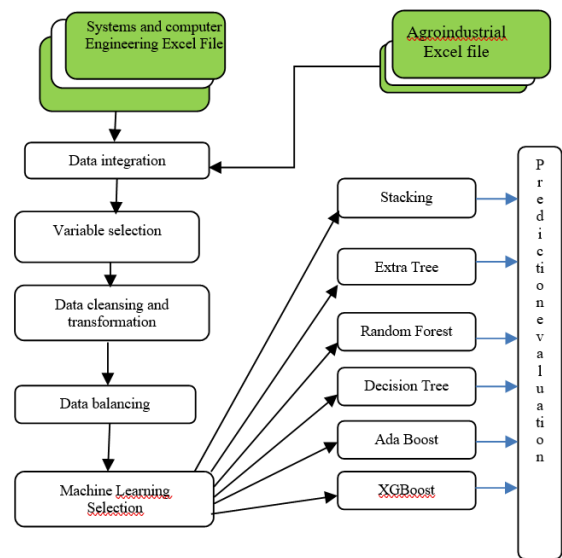


Fig. 2. Process of Prediction of Academic Performance through the Stacking Model.

TABLE I. LIST OF FEATURES USED IN THE QUESTIONNAIRE

Feature	Description	Feature	Description
SCHOOL	Academic school	ESTATUS_lab_padre	Employment status of the father
Cod_student	Student Code	ESTATUS_lab_madre	Mother's employment status
SEX	Gender	TYPE_viv	Type of housing
CI_ante	Previous cycle	INCOME_pa	Father's income
AVERAGE_acu_	Academic performance	INCOME_ma	Income of the mother
Prom_trans	Previous average	PLACE_res	Place of residence
NATIONALITY	Nationality	scholarship	Do you have a scholarship?
CURRENT_AGE	Current age	c_otra_carr	Do you have another career?
Anio_ingreso	Year of admission to the University	c_title_otra	Do you have another title?
AGE_estudiar	Age at which he began to study	DEPARTMENT	Department
FECH_nac	Date of birth	PROVINCE	Province
RACE	Race	DISTRICT	District
n_int_fami	Number of members in the household		

TABLE II. LIST OF FEATURES WITH THE HIGHEST CORRELATION

Feature	Domain
SEX	Nominal (Female, Male)
CURRENT_AGE	Whole
Anio_ingreso	Whole
AGE_estudiar	Whole
N_int_fami	Whole
ESTATUS_lab_padre	Nominal (Dependent, Independent)
ESTATUS_lab_madre	Nominal (Dependent, Independent)
INCOME_pa	Real
INCOME_ma	Real
AVERAGE_acu	Real

TABLE III. LIST OF FINAL FEATURE

Feature	Domain
SEX	Nominal (Female, Male)
CURRENT_AGE	Whole
Anio_ingreso	Whole
AGE_estudiar	Whole
N_int_fami	Whole
ESTATUS_lab_padre	Nominal (Dependent, Independent)
ESTATUS_lab_madre	Nominal (Dependent, Independent)
INCOME_pa	Real
INCOME_ma	Real
CONDITION	Nominal (Bad, Regular, Good)

d) *Data balancing*: Fig. 3 shows that the dataset is unbalanced, so accuracy could be affected. There is a lot of difference between the minority class (Good) and the majority class (Regular), so it was necessary to balance the dataset to ensure a better percentage of accuracy. The technique of oversampling has been used for data balancing as shown in Fig. 4, which generates artificial examples of the minority class, until reaching the number of records of the majority class.

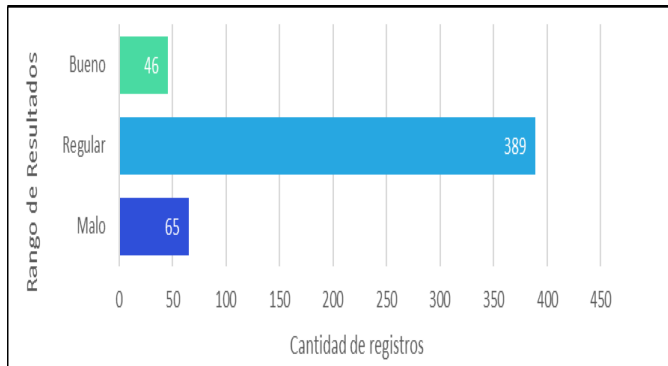


Fig. 3. Output Class Records before Applying Data Balancing.

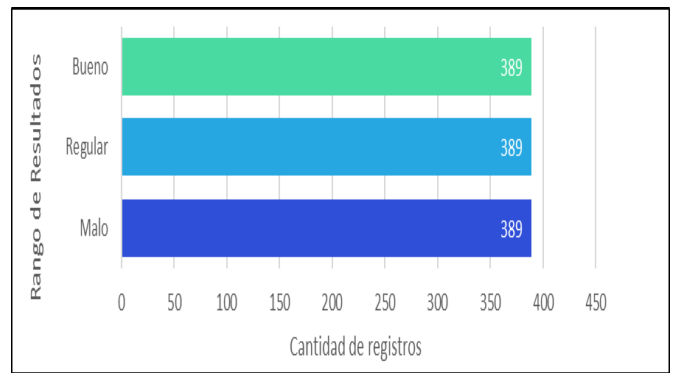


Fig. 4. Output Class Records after Applying Data Balancing.

4) *Transformation*: All development and implementation of the predictive model and data processing was done using Python software.

Due to the use of this tool, it was necessary for the entire dataset to have numeric values. That is why the data of the characteristics SEX, ESTATUS\_lab\_father, ESTATUS\_lab\_mother and CONDITION, had to be transformed into numbers according to conditions. Table IV shows the new values of the characteristics that were transformed.

5) *Data mining*: Python is a tool for data mining that offers many modules or libraries to be used professionally. These modules help the application of various classifiers.

For the application of data mining, one of the ways to increase the accuracy of the prediction made by a specific classification technique is the use of ensemble learning algorithms, one of them being stacking, which is what was applied for the construction of this predictive model.

According to the definition of [24] "it's the process of using different machine learning models one after another, where the predictions of each model are aggregated to create a new feature."

The techniques used as a basis for using stacking were: Extra trees, Random Forest, Decision Tree, Ada Boost, XGBoost.

TABLE IV. LIST OF FEATURES WITH NUMERIC VALUES

Feature	Numeric values
SEX	1 = Female 2 = Male
ESTATUS_lab_father	1 = Dependent 2 = Independent 3 = Deceased 4 = Live without a father
ESTATUS_lab_mother	1 = Dependent 2 = Independent 3 = Housewife 4 = Not working
CONDITION	1 = Bad 2 = Regular 3 = Good

The dataset was classified into two groups. The first group of data is for training and is made up of 2/3 of the total data. The second group is the test group and is made up of the remaining 1/3 of data. The application of the stacking algorithm was done using the training data to prepare the model, and then the test data. Model performance can be observed after application of the model to the test dataset.

6) *Interpretation and evaluation:* Python is used to import the dataset from an excel spreadsheet. The attributes with the highest correlation were selected to apply model training. The attributes sex, father's employment status, and mother's employment status were converted to numerical to avoid errors during the modeling process. Applied the different techniques to the final data set, each of these obtained different precisions.

To know the efficiency of the classifiers, these were evaluated using a confusion matrix, in which the number of records classified correctly and incorrectly is appreciated. 5 models were built to individually analyze the performance of the models and each of these obtained the following matrix confusion:

a) *Extra trees:* The first model to be built was the model based on the Extra Trees sorter. For its application in the Python tool, the ExtraTreesClassifier class of the sklearn.ensemble library was used and the following parameters were considered:

- random\_state = 0
- n\_jobs = -1
- n\_estimators = 100
- max\_depth = 3

Applied the algorithm to the dataset, the confusion matrix obtained was as follows:

As shown in Table V, is classified most students in the Bad category. These results show that the algorithm does not give good accuracy for this case.

b) *Random Forest:* For the implementation of this model, the RandomForestClassifier class of the sklearn.ensemble library was used and the parameters considered for its application were the following:

- random\_state = 0
- n\_jobs = -1
- n\_estimators = 100
- max\_depth = 3

TABLE V. CONFUSION MATRIX FOR THE EXTRA TREES MODEL

		Prediction		
		Bad	Regular	Well
Current	Bad	108	4	2
	Regular	96	16	22
	Well	97	2	42

The confusion matrix that was obtained from this model was as follows:

Based on the data obtained in Table VI, it can be said that the model has predicted well in terms of the classification of students in the Bad and Good categories, but for the Regular category, there is some uncertainty when classifying almost the same number of students for the three categories.

c) *Decision Tree:* To implement this algorithm, the DecisionTreeClassifier class of the sklearn.tree library was used and the parameters considered were the following:

- random\_state = 0
- min\_samples\_split = 2
- max\_depth = None

Running this model gets the following confusion matrix:

According to the values of the matrix, this is the model that has obtained a better precision, because as shown in Table VII, of 114 students correctly predicted the 114 within the Bad category, of 134 students correctly predicted 96 within the Regular category, and of the rest, 15 incorrectly predicted as Bad and 23 as Good. Finally, he correctly predicted 141 students as Good.

d) *Ada Boost:* This algorithm was implemented using the AdaBoostClassifier class of the sklearn.ensemble library and the parameters considered for its application were the following:

- random\_state = 0
- n\_estimators = 100

The confusion matrix that was obtained from this model was as follows:

Table VIII shown that of each dataset belonging to the three categories, the model was able to classify half of the students correctly, but not enough to be considered a good prediction, since it misclassified a large percentage of the students. So it is assumed that accuracy is not good for this model.

TABLE VI. CONFUSION MATRIX FOR THE RANDOM FOREST MODEL

		Prediction		
		Bad	Regular	Well
Current	Bad	87	9	18
	Regular	42	51	41
	Well	34	12	95

TABLE VII. CONFUSION MATRIX FOR THE DECISION TREE

		Prediction		
		Bad	Regular	Well
Current	Bad	114	0	0
	Regular	15	96	23
	Well	0	0	141

TABLE VIII. CONFUSION MATRIX FOR THE ADA BOOST MODEL

		Prediction		
		Bad	Regular	Well
Current	Bad	62	43	9
	Regular	27	77	30
	Well	9	45	87

e) *XGBoost*: For the implementation of this model, the *XGBClassifier* class of the *xgboost* library was used and the parameters considered for its application were the following:

- `random_state = 0`
- `n_jobs = -1`
- `learning_rate = 0.1`
- `n_estimators = 100`
- `max_depth = 3`

The confusion matrix obtained from this model was as follows:

This algorithm is the second most accurately after the decision tree. Table IX shown that of 114 students who belonged to the Bad category, it correctly ranked 108. Out of 141 students in the Good category, he correctly predicted 132. But of the 134 students considered regular, only 80 could predict correctly.

TABLE IX. CONFUSION MATRIX FOR THE XGBOOST MODEL

		Prediction		
		Bad	Regular	Well
Current	Bad	108	1	5
	Regular	23	80	31
	Well	0	9	132

What is part of a confusion matrix are the following four classifiers:

- True Positives (TP): These are the records correctly classified in the positive class.
- False Positives (FP): These are the records incorrectly classified in the positive class.
- False Negatives (FN): These are the records incorrectly classified in the negative class.
- True Negatives (TN): These are the records classified correctly in the negative class.

From these values, the following metrics can be calculated to evaluate the effectiveness of a predictive model:

- Sensitivity

This metric measures the positive values, in this case, correctly identifying students in the Bad, Regular and Good categories, according to the given parameters.

$$Sensitivity(TPR) = \frac{TP}{TP + FN}$$

- Specificity

This metric measures the negative or false values.

$$Specificity(TNR) = \frac{TN}{TN + FP}$$

- Precision

This metric measures the total number of items correctly classified as positive.

$$Precision(P) = \frac{TP}{TP + FP}$$

- Accuracy

$$Accuracy(ACC) = \frac{TP + TN}{TP + FP + FN + TN}$$

This metric measures the veracity of the prediction, that is, the difference between the predicted value and the actual one.

According to the results obtained from the application of the techniques individually, each of these obtained the following percentages in their precisions, shown in Table X:

TABLE X. METRIC RESULTS FOR INDIVIDUAL TECHNIQUES

Classification Technique	Metric			
	Sensitivity	Specificity	Precision	Accuracy
Extra trees	.4548	.7295	.5741	.6178
Random Forest	.6058	.8011	.6196	.7326
Decision Tree	.9054	.9509	.9144	.9348
Ada Boost	.5785	.7889	.5965	.7206
XGBoost	.8268	.9106	.8330	.8817

And after the application of stacking, the results he obtained were shown below in Table XI:

TABLE XI. METRICS RESULTS FOR STACKING

Classification Technique	Metric			
	Sensitivity	Especificidad	Sensitivity	Exactitud
Stacking	.9253	.9619	.9286	.9485

7) *Knowledge*: A graphical interface was created (Fig. 5) in which the previously developed predictive model was integrated, for the ease of use of teachers and in this way the performance of new engineering students can be predicted.

As a test, the following values were entered into the interface:

Age:

Sex:

Members:

And the result obtained concludes that the student will obtain a low academic performance.



Sexo  
 Femenino  
 Masculino

Edad Actual  
Ingresa tu edad actual

Año de Ingreso  
Ingresa el año en que ingresaste a la UNIS

Edad de Ingreso  
Ingresa la edad que tuviste cuando ingresaste a la UNIS

N° de integrantes  
Ingresa el número de integrantes en tu familia

Estatus laboral del padre  
Seleccionar

Estatus laboral de la madre  
Seleccionar

Ingreso de madre  
Ingresa el ingreso de tu madre

Ingreso de padre  
Ingresa el ingreso de tu madre

Predecir

© Copyright 2021, UNIS

Fig. 5. Web Interface.

#### IV. FINDINGS AND DISCUSSION

In this paper, the stacking technique was used to predict students' academic performance. In addition, as part of this technique, other classification methods such as Extra Trees, Random Forest, Decision Tree, AdaBoost and XGBoost were necessary, which are algorithms that obtain good results for a certain type of situation. During the modeling process, a cross-validation of 10 was required to ensure that each result is independent of division for training and test data.

The results obtained from the stacking technique (which is a combination of five algorithms) is 92.86% accuracy which is very encouraging with respect to the other results obtained by Extra Tree with 57.41% accuracy, Random Forest 61.96%, Decision Tree 91.44%, Ada Boost 59.65% and XGBoost with an accuracy of 83.3%, while authors such as [15], in their study show that the Naives bayes algorithm gave as a significant useful result an accuracy of 84%, being considered the best algorithm to predict academic performance and thus be able to arrive at solutions to improve the problem. On the other hand, in the study of the researchers [9], they developed a predictive model using the Rep Tree technique, which obtained an accuracy of 60.9%.

With regard to sensitivity, it is so that in the present study the Extra Trees technique was used, which reached a sensitivity of 45.48%, the Random Forest technique, 60.58%, the Decision Tree technique 90.54%, Ada Boost 57.85% and

finally XGBoost obtained a sensitivity of 82.68%, however when combining the aforementioned techniques through the Stacking technique a sensitivity of 92.53% was obtained, exceeding the percentage of sensitivity of the models; this is corroborated with the study of [16], which using the Naive Bayes technique, the model obtained a sensitivity of 66.7%.; they also [6] did a study in which a sensitivity of 88.7% was reached using Random Forest.

Likewise, with regard to specificity using the extra trees, Random Forest, Decision Tree, Ada Boost, XGBoost and each of these techniques, a specificity of 72.59%, 80.11%, 95.09%, 78.89%, 91.06% respectively was obtained and when performing the combination through the Stacking technique a specificity of 96.19% was obtained. In a study conducted by [14] they made use of the Naive Bayes technique, which obtained a specificity of 100%.

A limitation of the present study is that it was carried out considering the total average of the academic cycle of the students and in addition, they only belonged to the engineering schools. Results may vary for other schools or if a specific course or subject is considered.

Universities can employ the generation of a predictive model through stacking to predict the results of students' academic performance by cycle. Since it was demonstrated that its use and application can achieve a better accuracy in the prediction. This will help improve academic performance, as it will allow corrective action to be taken in advance and will also help reduce the percentage of students suffering from an academic delay.

#### V. CONCLUSION

The objective of this work is to develop a model that achieves better accuracy compared to the individual application of various techniques, through the stacking of different classification techniques, and in this way check if a better prediction is obtained.

The questionnaire applied to students to obtain the data set contains many questions, some of which have a greater correlation with academic performance than others. So maintaining the most important features benefits the accuracy of the prediction.

To achieve a good percentage of accuracy, it was necessary to have a process that involved cleaning the data, eliminating attributes less correlated with the output class, eliminating incomplete records or with invalid values. In this model, 5 classification techniques are used that are part of the multiclassification technique, stacking. The individual techniques were Extra Trees, Random Forest, Decision Tree, AdaBoost and XGBoost, after obtaining the predictions of each technique, XGBoost was used as a second-level technique to make the final prediction.

After having applied stacking it can be concluded that this has given better results than the application of the techniques individually.

#### REFERENCES

- [1] D. Ha, C. Giap and N. Huong, "An empirical study for student academic performance prediction using machine learning techniques",

- International Journal of Computer Science and Information Security.Vietnam, vol.18, no. 3, pp. 21-28, March 2020.
- [2] D. Rodríguez and R. Guzmán, “Rendimiento académico y factores sociofamiliares de riesgo. Variables personales que moderan su influencia”, *Perfiles educativos*. Perú, vol.41, no.164, pp.118-134 , Abril 2019.
- [3] M. Flores, H. Rivera and F. Sánchez, “Bajo rendimiento académico : más allá de los factores sociopsicopedagógicos”, *Revista Digital Internacional de Psicología y Ciencia Social. México*, vol. 2, no.1, pp. 95-104, Enero 2016.
- [4] S. Vadivukkarasi and S. Santhi. “A novel hybrid learning based Ada Boost (HLBAB) classifier for channel state estimation in cognitive networks”. *International Journal of Dynamics and Control*. India, pp. , 299-307, April 2021.
- [5] B. Pavlyshenko, “Using Stacking Approaches for Machine Learning Models”, 2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP).Ukraine, pp. 25-28, August 2018.
- [6] D. Aggarwal, S. Mittal, and V. Bali, “Prediction Model for Classifying Students Based on Performance using Machine Learning Techniques”, *International Journal of Recent Technology and Engineering (IJRTE)*. Indian, vol. 8, no. 257, pp. 496-503, July 2019.
- [7] S. Bhutto, I. Farah, Q. Ali and M. Anwar, “Predicting Students' Academic Performance Through Supervised Machine Learning” , 2020 International Conference on Information Science and Communication Technology (ICISCT). Pakistan, pp. 1-6, February 2020.
- [8] S. Nageswari, M. Pallavi and P. Divya, “Comparison of classification techniques on data mining”, *International Journal of Emerging Technology and Innovative Engineering*.India, vol.5, no.5, pp. 267-272, April 2019.
- [9] A. Hamoud, A. Hashim and W. Awadh, “Predicting student performance in higher education institutions using decision tree analysis”, *International Journal of Interactive Multimedia and Artificial Intelligence*. Iraq , vol. 5,no.2 , pp. 26-31, February 2018.
- [10] OEI. “Indicadores Socio-Económicos de los estudiantes de pregrado de la UNS”, *Nuevo Chimbote*: Universidad Nacional del Santa, 2019.
- [11] K. Rawat and I. Malhan, “A Hybrid Classification Method Based on Machine Learning Classifiers to Predict Performance in Educational Data Mining”, *Proceedings of 2nd International Conference on Communication, Computing and Networking*. Switzerland, vol.46, no.1., pp. 277-68, September 2019.
- [12] N. Chauhan, K. Shah, D. Karn and J. Dalal, “Prediction of Student's Performance Using Machine Learning”, 2nd International Conference on Advances in Science & Technology (ICAST). India, pp.1-5, April 2019.
- [13] S. Ahmad, S. Mutalib, H. Abdul and S. Abdul, “A Review on Student Attrition in Higher Education Using Big Data Analytics and Data Mining Techniques”, *I.J. Modern Education and Computer Science*. Malaysia, vol. 11, no. 8, pp. 1-14, August 2019.
- [14] Y. Widyaningsih, N. Fitriani and D. Sarwinda, “A Semi-Supervised Learning Approach for Predicting Student's Performance: First-Year Students Case Study”, 2019 12th International Conference on Information & Communication Technology and System (ICTS). Indonesia, pp. 291-295. July 2019.
- [15] I. Khan, A. Al Sadiri, A. Ahmad and N. Jabeur, “Tracking Student Performance in Introductory Programming by Means of Machine Learning”, 2019 4th MEC International Conference on Big Data and Smart City (ICBDSC). Malaysia,pp.1-6, January 2019.
- [16] A. Hamoud, A. Humandi, A. Awadh, and A. Hashim, “Students' success prediction based on Bayes algorithms”, *International Journal of Computer Applications*. Iraq, vol.178, no.7, pp.6-12, November 2017.
- [17] A. Ortiz, “Algoritmo multclasificador con aprendizaje incremental que manipula cambios de conceptos”, Granada: Universidad de Granada, 2014.
- [18] M. Camana, S. Ahmed, C. Garcia and I. Koo, “Extremely Randomized Trees-Based Scheme for Stealthy Cyber-Attack Detection in Smart Grid Networks”, *IEEE Access*. Korea, vol. 8, no.1 ,pp. 19921-19933, January 2020.
- [19] Y. Xiang, L. Li and W. Zhou, “Random Forest Classifier for Hardware Trojan Detection”, 12th International Symposium on Computational Intelligence and Design. China, pp. 134-137, December 2019.
- [20] E. Irfiani, I. Elyana, F. Indriyani, F. Schaduw and D. Harmoko, “Predicting Grade Promotion Using Decision Tree and Naïve Bayes Classification Algorithms”, 2018 Third International Conference on Informatics and Computing (ICIC). Indonesia, pp.1-4, October 2018.
- [21] Z.Yong, L. Jianyang, L. Hui and G. Xuehui, “Fatigue Driving Detection with Modified Ada-Boost and Fuzzy Algorithm”, 2018 Chinese Control And Decision Conference (CCDC).China,pp. 5971-5974, June 2018.
- [22] C.Wang, C. Denga and S. Wang, “Imbalance-XGBoost: leveraging weighted and focal losses for binary label-imbalanced classification with XGBoost”, *Pattern Recognition Letters*. China, vol. 136, no.1, pp. 190-197, August 2020.
- [23] S. Asante, P. Ngare, and D. Ikpe,“On Stock Market Movement Prediction Via Stacking Ensemble Learning Method”, 2019 IEEE Conference on Computational Intelligence for Financial Engineering & Economics (CIFER). Kenya, pp.1-8, May 2019.
- [24] IBM. (17. Enero 2020). Stack machine learning models: Get better results. Von IBM Developer: <https://developer.ibm.com/articles/stack-machine-learning-models-get-better-results/> abgerufen.

# COVID-19 Disease Detection based on X-Ray Image Classification using CNN with GEV Activation Function

Karim Ali Mohamed, Emad Elsamahy, Ahmed Salem  
College of Computing and Information Technology, Arab Academy for Science  
Technology and Maritime Transport (AASTMT), Cairo, Egypt

**Abstract**—The globe was rocked by unprecedented levels of disruption, which had devastating effects on daily life, global health, and global economy. Since the COVID-19 epidemic started, methods for delivering accurate diagnoses for multi-category classification have been proposed in this work (COVID vs. normal vs. pneumonia). XceptionNet and Dense Net, two transfer learning pre-trained model networks, are employed in our CNN model. The low-level properties of the two DCNN structures were combined and used to a classifier for the final prediction. To get better results with unbalanced data, we used the GEV activation function (generalized extreme value) to augment the training dataset using data augmentation for validation accuracy, which allowed us to increase the training dataset while still maintaining validation accuracy with the output classifier. The model has been put through its paces in two distinct scenarios. In the first instance, the model was tested using Image Augmentation for train data and the GEV (generalized extreme value) function for output class, and it got a 94% accuracy second instance Model evaluations were conducted without data augmentation and yielded an accuracy rating of 95% for the output class.

**Keywords**—COVID-19; CNN; GEV function; image augmentation

## I. INTRODUCTION

In December 2019 less than four months after coming in Wuhan for the first time, the coronavirus has deteriorated into a public health calamity. As of March 30, 2021, there were 127.34 million reported diseases and roughly 2.78 million deaths on a global scale [1]. The illness COVID-19 is brought on by a virus source that irritates the lungs and makes patients develop pneumonia. These pneumonia cases are treated and medicated very differently from those caused by other viruses or bacteria. In addition to the diagnosis, specific preventative measures are implemented if a person exhibits COVID-19 signs. To prevent the infection from spreading, the COVID-19 patient is isolated for a predetermined number of days. Therefore, it is crucial to accurately and promptly identify COVID-19-related pneumonia in order to stop the virus's transmission. In the medical field, machine learning algorithms for automated diagnosis have recently gained traction as a tool for physicians [2], [3]. Deep learning algorithms have been used to correctly classify skin cancer [4], [5], Breast cancer diagnosis [6], [7]. Psychiatric disorder classification, detection of pneumonitis using chest x-rays, and image segmentation, COVID-19 is most commonly diagnosed

RT-PCR is the method used here. The early detection and treatment of this condition necessitates the use of chest CT and X-ray imaging. It's still possible to find symptoms on CT scans, even with test results that come back negative[8], because the RT-sensitivity of PCRs has dropped to 70% from 60% before [9], [10]. A good approach for diagnosing COVID-19 pneumonia when paired with CT has been demonstrated [11]. CT scans are often clear for the first two days or so after symptoms start to appear. When COVID-19 pneumonia survivors underwent CT lung examinations 10 days following the onset of symptoms, the most substantial pulmonary pathology was found [12], [13]. This is a black and white picture of the body's internal organs. The X-ray is a medical diagnostic tool that has been around for a long time and is still commonly used today. An X-ray image of the thoracic cavity can detect chest infections and other lung illnesses including pneumonia, making X-ray imaging a viable alternative diagnostic method for COVID-19, in light of the present healthcare crises throughout the world. In order to create a COVID-19 case identification system based on machine learning, we specified the following goals.

- Helping radiologists and other medical professionals identify minute, slow changes in X-rays that could otherwise go undetected.
- Because radiologists are so expensive, many people in developing nations do not have access to them. They might use this technology to identify their X-ray images as pneumonia, COVID-19, or normal; to build a model to scan complex data like CT and MRI scans for COVID-19 cases.

## II. RELATED WORKS

Classification and recognition jobs have been proven to be an effective machine learning method. Different deep-learning techniques have been used by researchers to detect COVID-19 in clinical pictures such chest CT scans and X-rays Alakwaa, Wafaa, Nassef, and Amr Badr [14]. For the detection of COVID-19, several of these radiological imaging techniques have recently gained popularity, The segmented CT images were initially fed straight into 3D CNNs for classification, but this proved to be insufficient. Instead, nodule candidates in the Kaggle CT scans were first identified using a modified U-Net trained on LUNA16 data (CT scans with tagged nodules). The U-Net nodule detection method had a high rate of false

positives, so regions of segmented CT scans of the lungs were used to feed 3D convolutional neural networks (CNNs) to determine whether the CT scan was positive or negative for lung cancer. These regions were where the U-Net output had identified the most likely nodule candidates. The test set accuracy produced by the 3D CNNs was 86.6%.

Hemdan Ezz, Marwa A. and Mohamed Esmail [15] using 25 confirmed positive COVID-19 instances on 50 chest X-rays serve as the study's validation data. Seven distinct deep convolutional neural network designs, including the updated (VGG19) and the second version of Google MobileNet, are included in the COVIDX-Net, experiments and evaluate the COVIDX-Net, 80–20% of X-ray pictures were used for the model's training and testing stages, respectively. With f1-scores of 0.89 and 0.91 for normal and COVID-19.[16] COVID-Net was developed by Alexander Wong, Linda, Wang and Zhong Lin ,COVID-19 identification is based on a deep model that achieved 93.2 percent accuracy rate in categorizing (COVID-19, normal and pneumonia). Tartaglione, Enzo, et al .[17] For COVID-19, we advocate a combination of Deep Learning and Transfer Learning, which has been the most thoroughly studied field of research these papers examine the extent to which COVID-19 identification may be improved by modifying popular deep models. Abdul Hafeez and Muhammad Farooq [18] are two such men. COVID-ResNet, Radiograph Detection, and more there are 5941 chest images in the COVIDx dataset, and the ResNet was trained with X-ray images of varying sizes and learning rates, resulting in an accuracy of 96.23 percent using COVID-ResNet. Tzani and Ioannis D .[19] used chest X-rays with different hyper parameter settings to test 5 different models for COVID-19 detection, all 1427 X-rays showed 224 people who had Covid-19 sickness, 700 people who had confirmed common pneumonia, and 504 people who were healthy. VGG-19, InceptionNet, MobileNetV2, XceptionNet, and Inception ResNetV2 are five conventional CNN designs that have been evaluated for the job of categorizing X-rays using various model parameters like the number of untrainable layers and the top layer neural network classifier settings achieve 96.78 %, 98.66 %, and 96.46 %, respectively, are the greatest levels of accuracy. Sultan Mahmud and Kh. Mustafizur Rahman [20] used Convolutional Neural Networks (CNNs) in a deep learning model that is proposed to automatically identify COVID-19 disease using CXR (Chest X-ray) pictures. Model was trained using 10293 X-ray pictures, 875 of which were from COVID-19 instances. The collection includes three separate types of tuples: pneumonia, COVID-19, and normal cases. The empirical results demonstrate that, while using a CNN with fewer layers than those works, the suggested model achieved 97% specificity, 96.3% accuracy, 96% precision, 96% sensitivity, and 96% F1-score, which are better than the works currently available. Junaid Latief and Asif Iqba [21] Exception's CoroNet deep neural network was used to identify and diagnose COVID-19 in chest x-ray images. ImageNet and a chest X-ray dataset generated by merging COVID-19 and other publically accessible X-ray pictures were used to train this. Accuracy was attained in the model with 98% recall and 93% accuracy in three of the COVID instances after conducting several tests (COVID vs. pneumonia vs. normal). Rajib Kumar and Sagar Deep [22] used CNN models to

identify COVID from chest X-ray images, the ensembles network is comprised of three CNN networks that have been trained. There was a 91.99 percent success rate for the NASNet, MobileNet, and DenseNet. Hasan K. Naji, Hayder K, Fatlawi,

Ammar J [23] implements classifiers using both ensemble classification algorithms (Adaptive Boosting and Adaptive Random Forest). The study of the data revealed a striking correlation between the patient's age, the presence of a chronic illness, and the rate of recovery. The experimental results show that adaptive boosting classifiers perform exceptionally well, reaching 99% accuracy, while adaptive random forest classifiers scored just 91% accuracy, Mahmoud B. Rokaya, [24]. The work emphasizes the value of bioscience in identifying recovered patients from mortalities. The decision trees (DT) could distinguish between recovered patients and mortalities with 94% accuracy even with little data. A shallow dense network attained a 75% accuracy rate. However, the net reached 99% accuracy when a 10-fold approach was used with the same data. They gathered the data for this study from King Faisal Hospital. Two parameters had the highest power to distinguish between recovered patients and mortalities, according to PCA analysis. When trained using only calcium and hemoglobin, the shallow net provides an accuracy of 92%.

Convolutional Neural Networks (CNNs) are used in a deep learning model suggested by Sohaib Asif, Ming Zhao, Fengxiao Tang, and Yusen Zhu [25] to automatically identify COVID-19 disease using CXR (Chest X-ray) images. They use a model to assess the performance of various pre-trained deep learning models (InceptionV3, Xception, MobileNetV2, NasNet and DenseNet201). Second, a lightweight shallow convolutional neural network (CNN) architecture with a low false-negative rate is developed for identifying X-ray pictures of a patient. The data set used in this study includes 2,541 chest X-rays from two separate public databases that have been confirmed as COVID-19 positive and healthy cases. The suggested model's performance is compared to those of pre-trained deep learning models. According to the results, the proposed shallow CNN has a maximum accuracy of 99.68% and more importantly sensitivity, specificity and AUC of 99.66%, 99.70% and 99.98%.

### III. MATERIAL AND METHODS

According to the guidelines in this section, chest X-rays should be classified as normal, pneumonia, or COVID-19. The issue with medical imaging is the lack of huge data sets. Because it isn't recommended to start from scratch and build a DCNN, the medical images can be categorized by using the features learned through a process called transfer learning [26]. The ensemble architecture suggested here will make sure that all of the descriptors needed for picture classification are there, so that the process can go smoothly. To get features from photos, a layer called "Filter" is used. These features are combined and then applied to a FC classification. We used features from two trained models as a starting point for the proposed model, with Global Average Pooling added to them. This layer is to cut down on feature length, which means that there will be fewer neurons in the last classifier input layer. This is good because it reduces the number of parameters,

which makes it less likely that the network will become over fit [27]. In this section, we will list the methods used in this model.

### A. Convolutional Neural Networks

A convolutional layer, which is made up of groups of kernels or filters, is the most important part of a CNN. During training, the layer parameters are learned. Filters are often smaller in size than the original image, and each filter constructs an activation map by combining with the image. The filter is moved over the image height and width, and at each spatial position we calculate the dot product between each filter element and the input. Fig. 1 depicts the convolution process. When a filter is applied to an image, the first layer of the activation map (shown in blue in Fig. 1) is formed via convolution with the image blue component. This method is repeated for each image element to create the activation map. Stacking the activation maps of each filter is used to enable convolutional layers to build their output volume along the depth dimension. The output of a neuron helps equalize each component of the activation map.

As the previous data led to the conclusion, the size of input image is equal to the size of the corresponding filter because each neuron is connected to a small local area of the input image. There are also factors that are shared by all neurons in an activation map. Because the convolutional layer has such a strong local connection, the network is forced to train filters that respond strongly to a specific portion of the input. The first convolutional layer looks for low-level characteristics like lines. The next convolutional layers look for high-level features like shapes and individual objects, as shown in Fig. 1.

### B. Data Augmentation

Accumulating fresh training data from previously collected data is known as "data augmentation". It is possible to improve photographs using simple image processing methods such as padding, cropping, rotation, and flipping. In order to train neural networks, these edited photographs are added to the original collection of photos, increasing the data set size. Data augmentation is used to artificially expand the training data set [28].

Imaging and labels are regularly altered in medical images to achieve this effect. Contrary to popular belief, data augmentation is a common practise in the training industry. It is simple to generate and decrease overfitting in CNNs using data augmentation and regression. COVID Image [29] is an excellent example of an image augmentation technique that uses only a little amount of training data to create modified copies of training data sets that belong to the same class as the original images.

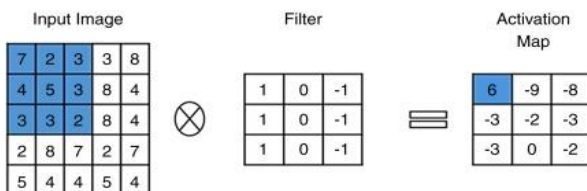


Fig. 1. Convolution Layer Sample.

The training data set is often utilized for data augmentation rather than validation or testing of the data sets. However, this is distinct from other data settings, such as pixel and image scaling. Consistency must be maintained across all datasets with which the model interacts. CNN models for deep learning have lately proved the necessity of data augmentation, since data augmentation enhances outcomes when training CNN on sparse data, but only when the augmentation procedures utilized are appropriate for each dataset [30].

There are a number of simple data augmentation strategies that are being tested. Table I shows the parameters utilized in the picture augmentation process.

### C. Transfer Learning Features Extraction and Concatenation

Each layer of the CNN learns ever-more-complex filters. The initial layers demonstrate how to employ fundamental feature detection filters, such as corners and edges, to look for things. They learn how to use filters to look for parts of things, like eyes and noses. This is how the last layers work. They develop the ability to recognize complete things in a variety of shapes and orientations. For the time being, I'll briefly describe what transfer learning is and how it works. How can you train an image classifier in a few hours? Training image classification models might take many days or even weeks, depending on the size of the networks and datasets used.

As a result, why do not we use the work of dedicated data scientists working on in image classification projects at businesses like Google and Microsoft as a starting point for the image classification initiatives? Transfer learning is based on the concept of taking pre-trained models, i.e. known-weights models, and applying them to a new machine learning issue. You cannot simply replicate the model and expect it to operate, you must retrain the network using the new data. However, because the weights from earlier layers are more generic, they can be frozen for training.

Consider pre-initialized networks to be intelligently constructed networks rather than a randomly generated network. Because we are effectively tailoring the network, lower learning rates are often employed in transfer learning than in regular network training. Transfer learning may not be beneficial if high learning rates are applied and the network's early layers are not frozen. In many transfer learning cases, just the final layer or a few layers are taught. There are many free neural networks available online that can be used for transfer learning if the problem is fairly general and the user doesn't have enough data to train the network — this is common.

TABLE I. IMAGE AUGMENTATION PARAMETER USED

Parameter	Value
samplewise_center	True
width_shift	0.23
height_shift	0.22
shear	0.15
zoom	0.15
horizontal_flip	Ture
brightness	[0.4,1.5]



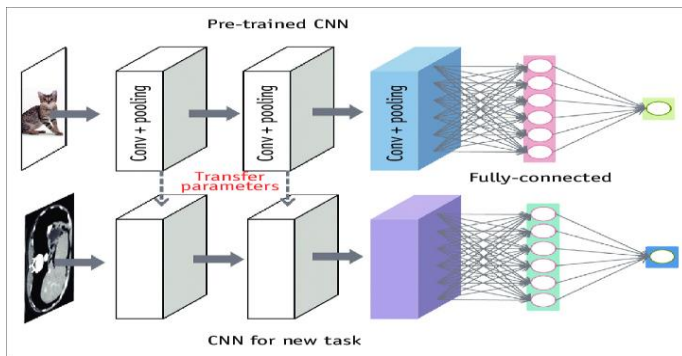


Fig. 2. Transfer Learning Model.

We may sum up this principle by saying that lower-level characteristics can be adapted to different environments by adjusting their weighting in later and fully connected layers, as we can see in Fig. 2, an example of a transfer learning model, as shown in Fig. 2.

The two-transfer model DCNN structures used in our model are described briefly below.

1) *DenseNet169*: In 2018, Huang et al. proposed densely linked convolutional networks, which interconnect each layer of a network in a feed-forward manner. A convolutional neural network with much more depth and accuracy was made feasible as a result of this groundbreaking achievement. Each layer of a dense network is linked directly to the next in a feed-forward method (inside every dense block). To create each subsequent layer, the feature mappings from prior layers are transferred to the new inputs.

2) *xception Net*: Deep detachable convolutions are used in the architecture of this CNN. A team of Google researchers came up with it. As a stepping stone between a regular convolution and a deep-separable convolution, Google has described the component units of convolutional neural networks. The input flow is the initial stop for data, followed by eight trips via the middle flow and a final stop at the outgoing flow. In addition, batch normalisation was used on all convolution and separable convolution layers in the final product.

#### D. GEV Activation Function

The majority of the data in a dataset is organized into a few number of classes, whereas several classes appear only sporadically. There is a considerable tail to the data in this example. Students who took classes with a larger number of students had a greater impact on their learnt traits. In this scenario, it's easier to simulate the more frequent classes than the rarer ones [31]. In both, binary and multiclass contexts, this problem exists.

When dealing with data that is very asymmetrical, with many instances in one class and few in the other, new techniques are needed. GEV distribution from extreme value theory yields a better activation function than sigmoid activation function when one class dominates the other. Binary and multiclass classification can be improved using

GEV activation functions rather than sigmoid or softmax activation functions. COVID-19 and other diseases with limited training examples may benefit from this new paradigm. When one side of the training dataset is much better than the other, or when the dataset is very imbalanced. CNN may then be used to extract the pictures' characteristics. The characteristics are then reduced to a single value by a liner combination in the fully linked layer. GEV (Generalized Extreme Value) is the activation function used to transform this single value into a probability [32]. GEV activation is provided by the function.

$$GEV(x|\mu, \sigma, \varepsilon) = \begin{cases} \exp\left\{-\exp\left(-\frac{x-\mu}{\sigma}\right)\right\}, & \text{if } \varepsilon = 0, \\ \exp\left\{-\left\{1 + \varepsilon\left(\frac{x-\mu}{\sigma}\right)\right\}^{-\frac{1}{\varepsilon}}\right\}, & \text{if } \varepsilon \neq 0, \end{cases} \quad (1)$$

Where  $\mu$ ,  $\sigma$ , and  $\xi$  are in the deep learning framework, parameters must be learned. According to the extreme value theorem the properly normalized maximum of a sample of independent and identically distributed random variables, can only converge to the GEV distribution. To give a probability, the GEV activation rescales the values between zero and one.

The parameters, on the other hand, enable the curve to better model the long-tail distribution that occurs with extreme data [33].

#### E. Evaluation of Performance

The most crucial metric for assessing how well our deep learning classifiers perform is accuracy. It's the sum of true positive and negative values divided by the total value

$$accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (2)$$

Precision is a measure of how many predictions in a certain class are really in that class.

$$Precision = \frac{TP}{TP+FP} \quad (3)$$

Recall is a metric that measures how many correct class predictions might be produced given all of the data that was found to be correct.

$$Recall = \frac{TP}{TP+FN} \quad (4)$$

The F1 score is measurement accuracy metric. The F1 score is equal to twice the ratio of the accuracy and recall measurements multiplied by the total of the accuracy and recall measures.

$$F1 \text{ score} = 2 * \frac{Precision * Recall}{Recall + precision} \quad (5)$$

#### F. Data Collection

The images were captured from a number of publically available resources.

- <https://www.kaggle.com/prashant268/chest-xray-covid19-pneumonia>
- <https://github.com/agchung>
- <https://github.com/ieee8023/covid-chestxray-dataset>



Three sub-volumes comprises each of the two volumes (“training and testing”) in which the data is organized and labelled into (COVID19, PNEUMONIA and NORMAL). Which contains 6,432 X-ray pictures make up the dataset divided as follows 1583 image as normal and 576 labelled COVID-19 and 4273 as pneumonia, which 20% of the data are test images. As seen in Fig. 3 and Fig. 4, we can see the dataset is highly unbalanced, we used GEV function to solve this problem. We plot some samples of data in Fig. 5.

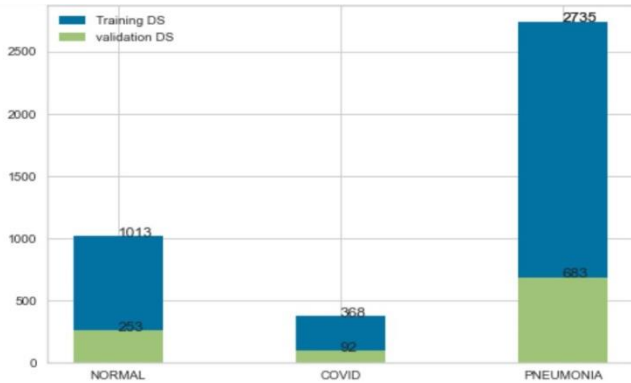


Fig. 3. Training and Validation Dataset.

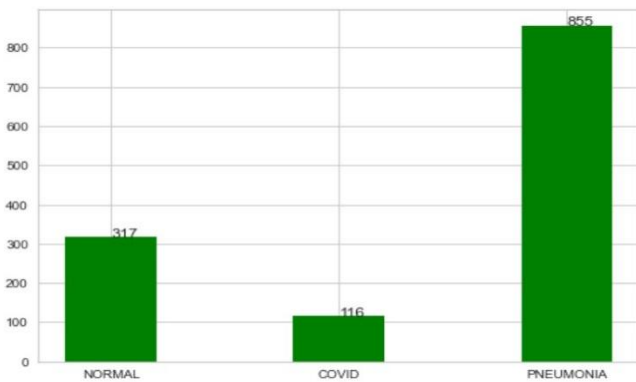


Fig. 4. Test Dataset.

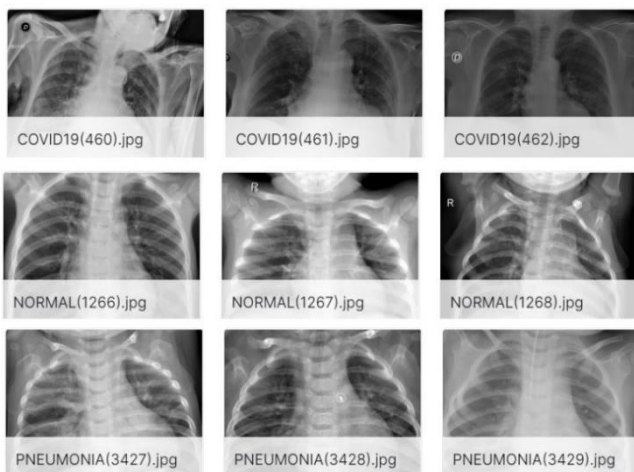


Fig. 5. Samples of Data.

#### IV. RESULTS AND DISCUSSION

COVID-19 detection is the primary goal of this investigation. We used a dataset that was organized into (COVID19, PNEUMONIA, and NORMAL) instances and used our model to classify the COVID-19 by employing features derived from two separate transfer learning models using GEV activation function to accomplish our study goal. In our models, we have a three-step process. Every step of the process will be described in depth, including pre-processing, categorization, and validation. Fig. 6 depicts the suggested framework. During the classification phase, we feed the proposed model the "feature maps" generated by two previously trained models (DensNet 169 and xception Net). Filters or feature detectors are applied to the input image, and the outcomes of those operations are calculated to produce these feature maps. Based on the model's design, a variety of feature maps was generated. Table II displays the feature maps generated by the DensNet 169 model, which were [7, 7, 1664].

Using the ImageNet dataset, the input of the 224x224x3 form is sampled down to 7x7x1664 at the conclusion of the model. Table III indicates the xception Net feature maps generated, which is [7, 7, 2048] (see Table III). 224x224x3 input is down sampled to 7x7x2048 at the model's conclusion. ImageNet-trained model structure was used to create this feature. Table IV's concatenation output feature that was generated from two preview Models [7, 7, 3712], is the input to our proposed model. In the form of a feature map created by merging two previously trained models, this layer significantly speeds up deep network training and boosts neural network robustness by normalizing data between neural network layers instead of normalizing raw data. Instead of analyzing the full dataset, a flattening layer reduces the result of normalization to a single-dimensional feature vector, which aids learning by speeding up training and increasing the pace at which information is absorbed. The flattening layer combines all the pixel data from convolutional layers into a single vector.

Once the model has received the vector, it uses it as an input layer. To feed data to each neuron in our model, we utilise the flatten function, which reduces multi-dimensional input tensors to just one dimension. Flattening layer output is shown in Table IV as [3712]. When the vector data has been flattened, it is transmitted to the CNN's layers, which are referred to as "completely linked" or "dense layers," where it is processed in one of two ways. Because of these interconnections between neurons at every level, the brain may function as a single unit. It is the initial responsibility of dense layers to categories the picture using the flattened output results of convolution and pooling layers as input.

As a result, the categorization determination is ultimately driven by the completely linked layer. We used three fully connected (FC) layers, which represent the global averaged attributes of the two models utilizing three neural layers, were used to solve the classification problem. There are 128 nodes, 32 nodes, and then 3 nodes in each layer of the brain. Every FC layer has a PReLU Activation Function applied. Neurons can be activated or deactivated using an activation function. It

will be determined whether or not the input from the neuron to the network is crucial, using simple mathematical approaches.

TABLE II. DENSENET 169 ARCHITECTURE

Layer	Output Shape	Parameter
Input Layer	224,224,3	0
DenseNet169	7,7,1664	12642880
Normalization	7,7,1664	6656
Global Average	1664	0
Flatten	1664	0
Dropout	1664	0
Dense	128	213120
PReLU	128	128
Dropout_1	128	0
Dense_1	32	4128
PReLU_1	32	32
Dropout_2	32	0
Dense_2	3	99

TABLE III. XCEPTION MODEL ARCHITECTURE

Layer	Output Shape	Parameter
Input Layer	224,224,3	0
XceptionNet	7,7,2048	20861480
Normalization	7,7,2048	8192
Global Average	2048	0
Flatten	2048	0
Dropout	2048	0
Dense	128	262272
PReLU	128	128
Dropout_1	128	0
Dense_1	32	4128
PReLU_1	32	32
Dropout_2	32	0
Dense_2	3	99

TABLE IV. PROPOSED MODEL ARCHITECTURE

Layer	Output Shape	Parameter
Input Layer	224,224,3	0
DenseNet169	7,7,1664	12642880
XceptionNet	7,7,2048	20861480
Concatenate	7,7,3712	0
Normalization	7,7,3712	14848
Global Average	3712	0
Flatten	3712	0
Dropout	3712	0
Dense	128	475264
PReLU	128	128
Dropout_1	128	0
Dense_1	32	4128
PReLU_1	32	32
Dropout_2	32	0
Dense_2	3	99
GEV	3	7

It is common for CNNs to include a "dropout layer," a mask that eliminates particular neurons from the next layer while keeping all others intact. It is possible to apply a dropout layer to an input vector in two ways: either to eliminate part of the vector's attributes, or to delete neurons inside a hidden layer. The value [0.5] of Dropout was utilized in the first FC layer in order to avoid CNNs being unduly dependent on the training data. This implies that 50% of the neurons in the input were randomly deactivated. This model relies heavily on the learning rate parameter. The rate at which we learn determines how frequently we need to adjust the settings we're working with. The model will take a long time to converge if the learning rate is too low, because the parameters will only change by modest amounts. If the learning rate is excessively high, the parameters may hop over the low spaces of the loss function, and the network may never achieve a convergent state. The inverse is also true. Learning Rate (0.0003 to 0.00005) was the range of the steps we utilized on this model schedule. Once the model has ceased improving, this callback method will attempt to change the model by decreasing the learning rate. Up to 13 epochs of training data were utilized with the validation data (validation loss').

Using a GEV function to forecast the output class improved performance in the imbalance class, when there are a lot of samples from one class and a few from the other, but only a few instances from the other. GEV distribution from extreme value theory yields a better activation function than sigmoid activation function when one class dominates the other. Binary and multiclass classification can be improved using GEV activation functions rather than sigmoid or softmax activation functions, as shown in Fig. 6.

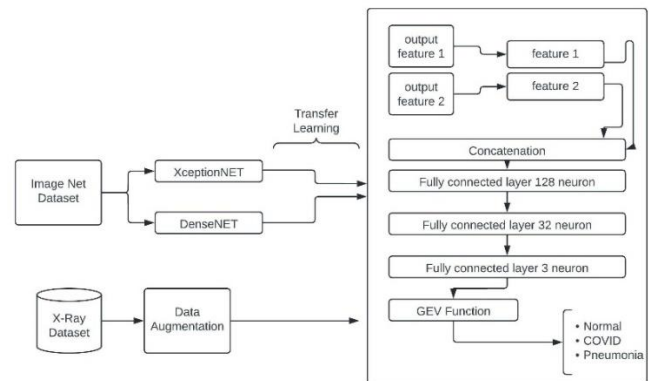


Fig. 6. Proposed Framework.

### A. Results and Evaluation Model

The effectiveness of the proposed model is also compared to that of Dense Net and xception Net as single classifiers. As illustrated in Table V, to compare model output, we repeat the model without the image augmentation parameter, as shown in Table VI.

When compared to two independent models, Xception-NET and DenseNet, without augmentation, the estimated model accuracy is 95.5%, and with augmentation, it is 94% as we can see in Fig. 11. As shown in Fig. 7 and 8, DenseNet achieves an accuracy of 93.2 percent with picture

augmentation and 94 percent without image augmentation, whereas Xception NET achieves an accuracy of 84 percent with image augmentation and 94 percent without image augmentation. 13 is the ideal number of epochs, as we can see in Fig. 12 based on intersection of the training line with validation line so we stopped the training model at epoch 13.

We can measure the metrics of the outcomes of our categorization investigation using the confusion matrix. The confusion matrix for the proposed CNN framework's test cases is shown in Fig. 13 with image Augmentation and Fig. 14 without, In addition, Fig. 9, 10 graphic representation of the CNN classifier performance evaluation shows loss both with and without image augmentation during the validation and training stages. Additionally, at epoch number 12, the validation and training losses attained by the suggested system are 0.1587 without image augmentation and 0.1777 with augmentation.

TABLE V. RESULT WITH IMAGE AUGMENTATION TECHNIQUE

Model Name	Precision	Recall	F1-score	Accuracy
Densnet	0.90	0.94	0.92	0.93
Xception	0.80	0.85	0.81	0.84
Proposed model	0.93	0.93	0.93	0.94

TABLE VI. RESULT WITHOUT IMAGE AUGMENTATION

Model Name	Precision	Recall	F1-score	Accuracy
Densnet	0.94	0.93	0.93	0.94
Xception	0.94	0.93	0.94	0.94
Proposed model	0.95	0.94	0.95	0.95

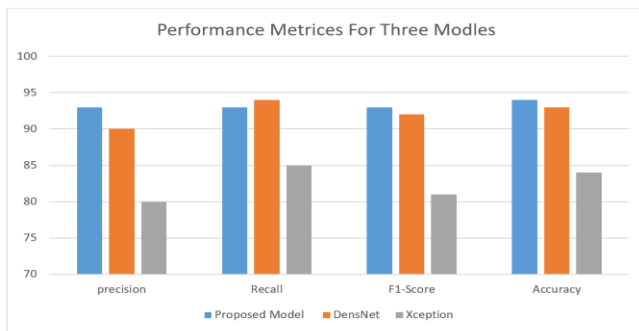


Fig. 7. Comparison of Three Model using Image Augmentation.

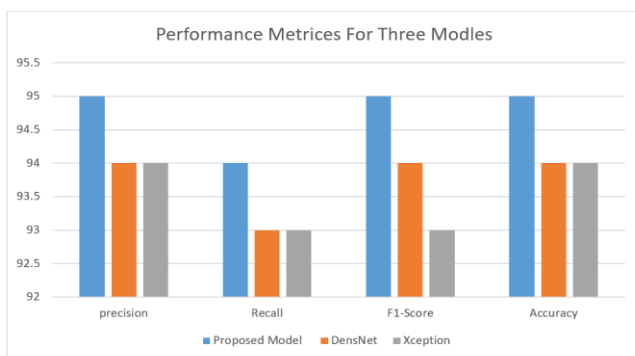


Fig. 8. Three Models are Compared without Image Augmentation.

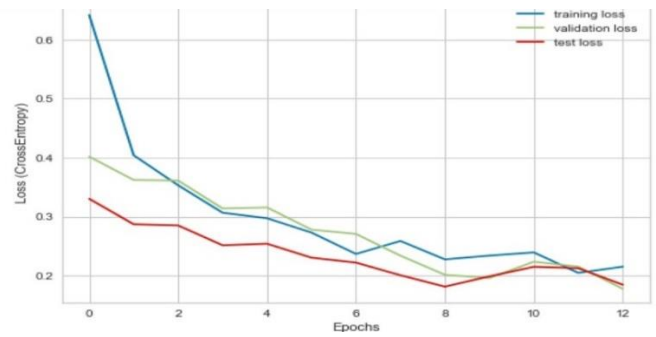


Fig. 9. Loss Value Plot with Image Augmentation.

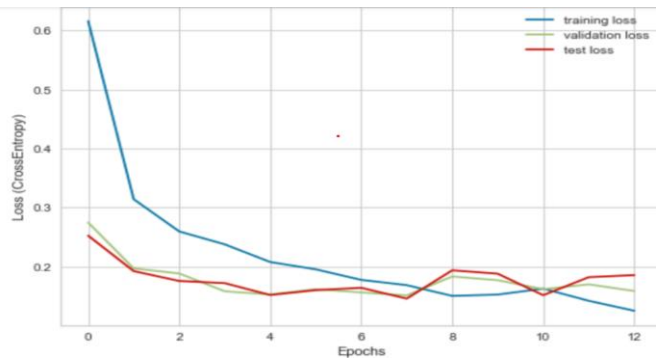


Fig. 10. Loss Value Plot without Augmentation.

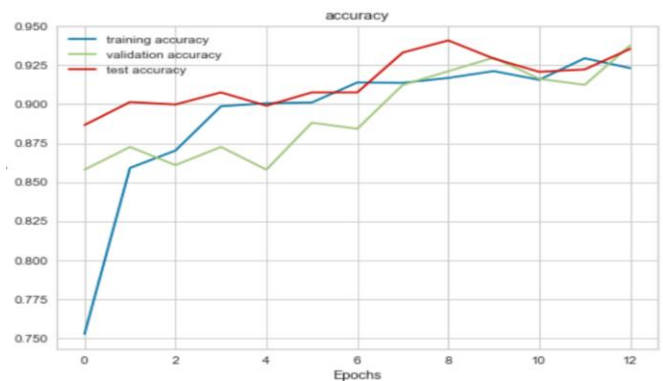


Fig. 11. Accuracy Graph of the Suggested Model's using Image Augmentation.

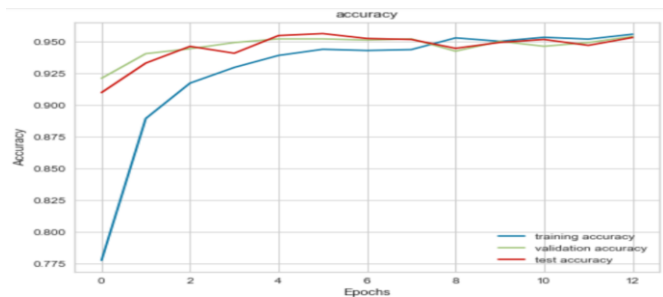


Fig. 12. Accuracy Graph of the Suggested Model's without Image Augmentation.

Confusion matrix

		Covid	Normal	pneumonia
Actual	Covid	108	2	6
	Normal	0	298	19
	pneumonia	1	49	805
		prediction		

Fig. 13. Confusion Matrix of Suggested Model's with Image Augmentation.

		Covid	Normal	pneumonia
Actual	Covid	113	0	4
	Normal	0	298	37
	pneumonia	3	19	814
		prediction		

Fig. 14. Confusion Matrix of Suggested Model's without Image Augmentation.

## V. CONCLUSION

Using the GEV activation function, For the purpose of identifying and classifying COVID-19 occurrences from X-ray images, we suggested a deep learning model using two DCNN structures. 95 percent of the time, our model is able to handle jobs that include numerous classes. To process the COVID dataset without data augmentation, our model achieved 95% accuracy in just 13 learning cycles. The GEV Function surpasses a single classifier in terms of generalization performance when features are combined from the two DCNN structures without picture augmentation. Radiologists can benefit from the suggested strategy by learning more about COVID-19's important components. Accuracy is expected to increase better with more and more training data. The following are some of the most important discoveries from this research: For effective and more accurate image categorization, CNN models require a sufficient number of images.

When employing an existing dataset with a GEV activation function, picture augmentation parameters have little impact on the performance of a CNN model.

In a statistically significant way, the suggested CNN model improves the performance of other single CNN models. The medical sector may greatly benefit from CNN-based diagnosis using X-ray imaging when dealing with large-scale testing situations like COVID-19.

## REFERENCES

[1] "WHO Coronavirus (COVID-19) Dashboard", Covid19.who.int, 2022. [Online]. Available: <https://covid19.who.int/>. [Accessed: 25- Sep-2022].

[2] G. Litjens et al., "A survey on deep learning in medical image analysis", *Medical Image Analysis*, vol. 42, pp. 60-88, 2017. Available: 10.1016/j.media.2017.07.005.

[3] J. Ker, L. Wang, J. Rao and T. Lim, "Deep Learning Applications in Medical Image Analysis", *IEEE Access*, vol. 6, pp. 9375-9389, 2018. Available: 10.1109/access.2017.2788044.

[4] A. Esteva et al., "Dermatologist-level classification of skin cancer with deep neural networks", *Nature*, vol. 542, no. 7639, pp. 115-118, 2017. Available: 10.1038/nature21056.

[5] N. Codella et al., "Deep learning ensembles for melanoma recognition in dermoscopy images", *IBM Journal of Research and Development*, vol. 61, no. 45, pp. 5:1-5:15, 2017. Available: 10.1147/jrd.2017.2708299.

[6] Y. Celik, M. Talo, O. Yildirim, M. Karabatak and U. Acharya, "Automated invasive ductal carcinoma detection based using deep transfer learning with whole-slide images", *Pattern Recognition Letters*, vol. 133, pp. 232-239, 2020. Available: 10.1016/j.patrec.2020.03.011.

[7] H. Wang et al., "Mitosis detection in breast cancer pathology images by combining handcrafted and convolutional neural network features", *Journal of Medical Imaging*, vol. 1, no. 3, p. 034003, 2014. Available: 10.1117/1.jmi.1.3.034003.

[8] Q. Chen, Z. Zu, M. Jiang, L. Lu, G. Lu and L. Zhang, "Infection Control and Management Strategy for COVID-19 in the Radiology Department: Focusing on Experiences from China", *Korean Journal of Radiology*, vol. 21, no. 7, p. 851, 2020. Available: 10.3348/kjr.2020.0342.

[9] J. Kanne, B. Little, J. Chung, B. Elicker and L. Ketai, "Essentials for Radiologists on COVID-19: An Update—Radiology Scientific Expert Panel", *Radiology*, vol. 296, no. 2, pp. E113-E114, 2020. Available: 10.1148/radiol.2020200527.

[10] X. Xie, Z. Zhong, W. Zhao, C. Zheng, F. Wang and J. Liu, "Chest CT for Typical Coronavirus Disease 2019 (COVID-19) Pneumonia: Relationship to Negative RT-PCR Testing", *Radiology*, vol. 296, no. 2, pp. E41-E45, 2020. Available: 10.1148/radiol.2020200343.

[11] E. Lee, M. Ng and P. Khong, "COVID-19 pneumonia: what has CT taught us?", *The Lancet Infectious Diseases*, vol. 20, no. 4, pp. 384-385, 2020. Available: 10.1016/s1473-3099(20)30134-1.

[12] A. Bernheim et al., "Chest CT Findings in Coronavirus Disease-19 (COVID-19): Relationship to Duration of Infection", *Radiology*, vol. 295, no. 3, p. 200463, 2020. Available: 10.1148/radiol.2020200463.

[13] F. Pan et al., "Time Course of Lung Changes at Chest CT during Recovery from Coronavirus Disease 2019 (COVID-19)", *Radiology*, vol. 295, no. 3, pp. 715-721, 2020. Available: 10.1148/radiol.2020200370.

[14] W. Alakwaa, M. Nassef and A. Badr, "Lung Cancer Detection and Classification with 3D Convolutional Neural Network (3D-CNN)", *International Journal of Advanced Computer Science and Applications*, vol. 8, no. 8, 2017. Available: 10.14569/ijacsa.2017.080853.

[15] M. Karar, E. Hemdan and M. Shouman, "Cascaded deep learning classifiers for computer-aided diagnosis of COVID-19 and pneumonia diseases in X-ray scans", *Complex & Intelligent Systems*, vol. 7, no. 1, pp. 235-247, 2020. Available: 10.1007/s40747-020-00199-4.

[16] L. Wang, Z. Lin and A. Wong, "COVID-Net: a tailored deep convolutional neural network design for detection of COVID-19 cases from chest X-ray images", *Scientific Reports*, vol. 10, no. 1, 2020. Available: 10.1038/s41598-020-76550-z.

[17] E. Tartaglione, C. Barbano, C. Berzovini, M. Calandri and M. Grangetto, "Unveiling COVID-19 from CHEST X-Ray with Deep Learning: A Hurdles Race with Small Data", *International Journal of Environmental Research and Public Health*, vol. 17, no. 18, p. 6933, 2020. Available: 10.3390/ijerph17186933.

[18] Farooq, M., & Hafeez, A. (2020). Covid-resnet: A deep learning framework for screening of covid19 from radiographs. arXiv preprint arXiv:2003.14395.

[19] I. Apostolopoulos and T. Mpesiana, "Covid-19: automatic detection from X-ray images utilizing transfer learning with convolutional neural networks", *Physical and Engineering Sciences in Medicine*, vol. 43, no. 2, pp. 635-640, 2020. Available: 10.1007/s13246-020-00865-4.

[20] M. Islam, G. Stea, S. Mahmud and K. Rahman, "COVID-19 Cases Detection from Chest X-Ray Images using CNN based Deep Learning

- Model", International Journal of Advanced Computer Science and Applications, vol. 13, no. 5, 2022. Available: 10.14569/ijacsa.2022.01305108 [Accessed 18 August 2022].
- [21] A. Khan, J. Shah and M. Bhat, "CoroNet: A deep neural network for detection and diagnosis of COVID-19 from chest x-ray images", Computer Methods and Programs in Biomedicine, vol. 196, p. 105581, 2020. Available: 10.1016/j.cmpb.2020.105581.
- [22] Deb, S. D., & Jha, R. K. (2020, December). Covid-19 detection from chest x-ray images using ensemble of cnn models. In 2020 International Conference on Power, Instrumentation, Control and Computing (PICC) (pp. 1-5). IEEE.
- [23] H. Naji, H. Fatlawi, A. Karkar, N. GOGA, A. Kiss and A. Al-Rawi, "Prediction of COVID-19 Patients Recovery using Ensemble Machine Learning and Vital Signs Data Collected by Novel Wearable Device", International Journal of Advanced Computer Science and Applications, vol. 13, no. 7, 2022. Available: 10.14569/ijacsa.2022.0130792 [Accessed 25 September 2022].
- [24] M. Rokaya, "Shallow Net for COVID-19 Classification Based on Biomarkers", International Journal of Advanced Computer Science and Applications, vol. 13, no. 6, 2022. Available: 10.14569/ijacsa.2022.0130613 [Accessed 25 September 2022].
- [25] S. Asif, M. Zhao, F. Tang and Y. Zhu, "A deep learning-based framework for detecting COVID-19 patients using chest X-rays", Multimedia Systems, vol. 28, no. 4, pp. 1495-1513, 2022. Available: 10.1007/s00530-022-00917-7.
- [26] U. Abubakar, M. Boukar and S. Adeshina, "Evaluation of Parameter Fine-Tuning with Transfer Learning for Osteoporosis Classification in Knee Radiograph", International Journal of Advanced Computer Science and Applications, vol. 13, no. 8, 2022. Available: 10.14569/ijacsa.2022.0130829.
- [27] Zhang, X., Zhang, Y., Han, E. Y., Jacobs, N., Han, Q., Wang, X., & Liu, J. (2018). Classification of whole mammogram and tomosynthesis images using deep convolutional neural networks. IEEE transactions on nanobioscience, 17(3), 237-242.
- [28] N. Aminuddin, Z. Tukiran, A. Joret, R. Tomari and M. Morsin, "An Improved Deep Learning Model of Chili Disease Recognition with Small Dataset", International Journal of Advanced Computer Science and Applications, vol. 13, no. 7, 2022. Available: 10.14569/ijacsa.2022.0130750.
- [29] S. Govathoti, A. Reddy, D. Kamidi, G. BalaKrishna, S. Padmanabhuni and P. Gera, "Data Augmentation Techniques on Chilly Plants to Classify Healthy and Bacterial Blight Disease Leaves", International Journal of Advanced Computer Science and Applications, vol. 13, no. 6, 2022. Available: 10.14569/ijacsa.2022.0130618.
- [30] Rodrigues, Larissa Ferreira, Murilo Coelho Naldi, and Joao Fernando Mari. "Comparing convolutional neural networks and preprocessing techniques for HEP-2 cell classification in immunofluorescence images." Computers in biology and medicine 116 (2020): 103542.
- [31] J. Johnson and T. Khoshgoftaar, "Survey on deep learning with class imbalance", Journal of Big Data, vol. 6, no. 1, 2019. Available: 10.1186/s40537-019-0192-5.
- [32] Bridge, J., Meng, Y., Zhao, Y., Du, Y., Zhao, M., Sun, R., & Zheng, Y. (2020). Introducing the GEV activation function for highly unbalanced data to develop COVID-19 diagnostic models. IEEE journal of Biomedical and Health Informatics, 24(10), 2776-2786.
- [33] Bridge, J. T., & Zheng, Y. (2021). mGEV: Extension of the GEV Activation to Multiclass Classification.



# Deep Learning based Cervical Cancer Classification and Segmentation from Pap Smears Images using an EfficientNet

Krishna Prasad Battula<sup>1</sup>  
Research scholar, School of CSE  
VIT-AP University, Amaravati  
Andhra Pradesh, India

Dr B. Sai Chandana<sup>2</sup>  
School of CSE, VIT-AP University  
Amaravati  
Andhra Pradesh, India

**Abstract**—One of the most prevalent cancers in the world, cervical cancer claims the lives of many people every year. Since early cancer diagnosis makes it easier for patients to use clinical applications, cancer research is crucial. The Pap smear is a useful tool for early cervical cancer detection, although the human error is always a risk. Additionally, the procedure is laborious and time-consuming. By automatically classifying cervical cancer from Pap smear images, the study's goal was to reduce the risk of misdiagnosis. For picture enhancement in this study, contrast local adaptive histogram equalization (CLAHE) was employed. Then, from this cervical image, features including wavelet, morphological features, and Grey Level Co-occurrence Matrix (GLCM) are extracted. An effective network trains and tests these derived features to distinguish between normal and abnormal cervical images by using EfficientNet. On the aberrant cervical picture, the SegNet method is used to identify and segment the cancer zone. Specificity, accuracy, positive predictive value, Sensitivity, and negative predictive value are all utilized to analyze the suggested cervical cancer detection system performances. When used on the Herlev benchmark Pap smear dataset, results demonstrate that the approach performs better than many of the existing algorithms.

**Keywords**—Cervical cancer; pap smear; time-consuming; contrast local adaptive histogram equalization (CLAHE); Grey Level Co-occurrence Matrix (GLCM); morphological features; wavelet; SegNet

## I. INTRODUCTION

A thin layer of tissues made up of cells covers the human cervix. Cervical cancer is the term used to describe a condition when a cell is transformed into a malignant cell that can divide and expand quickly to form a tumor [1]. Cervical cancer, which affects women worldwide and ranks as the second leading cause of cancer-related mortality, is fatal. If this cancer is found early enough, it may be treated [2]. Typically, a biopsy and screening procedure is used to make the diagnosis. Techniques for image processing can be used to determine where cancer has spread. The fourth-most frequent cancer-related cause of death in women is cervical cancer [3, 4].

Intelligent systems and medical image processing both contribute to the analysis of cancerous cells. They grow more time and money efficient as new approaches are developed [5-7]. They are currently gaining popularity in place of traditional

techniques including Pap smears, colposcopies, and Cervicography [8]. These methods are objective to the human experience, but it's important to note that they don't completely replace the professional doctor's subjective assessment, even though they can greatly aid it [9].

Analyses of the nucleus and cytoplasm are typically necessary for cell classification investigations to take cell type into account. Consequently, it is essential to develop algorithms that would aid in nuclei and cytoplasm segmentation [10-12]. The majority of feature extraction uses the same standards that specialists use to evaluate a cell. However, there is limited knowledge of cervical cytology [13]. Although this has not yet been researched, the cell might possess traits found in higher features. As a result, deep learning techniques have recently made representational learning more well-known. The automatic extraction of characteristics from input photos is a remarkable benefit of deep learning [14].

Analyses of the nucleus and cytoplasm are typically necessary for cell classification investigations to determine the type of cell. As a result, algorithms that can divide the cytoplasm and nuclei into separate parts must be developed [15-17]. The majority of feature extraction considers cells using the same standards as experts. Cervical cytology is, however, not well understood. The cell might possess traits found in higher forms, however, this has not yet been researched [18].

As a result, deep learning techniques have recently made representational learning more well-known. The automatic extraction of characteristics from input photos is a remarkable benefit of deep learning. As a result, automatic screening has made extensive use of deep learning [19]. In particular, DeepPap achieved a predictive performance of 98.6% on the Herlev benchmark Pap smear dataset and comparable effectiveness on the HEMLBC private Pap smear dataset using patch extraction from the nucleus ground truth mask and a transfer learning strategy to initialize weights with a pre-trained model.

The main cause of cervical cancer-related deaths in female patients is that the disease cannot be identified at an earlier stage, and patients do not experience any symptoms until cancer has progressed to its terminal stage [20]. Only if it is



discovered at an earlier stage can the death ratio for female patients be decreased. To prevent patient deaths, this research suggests a mechanism for detecting cervical cancer at an earlier stage.

The major key contributions of the research are as follows,

- The preprocessing stages are used to increase the classification efficiency, even more, here images can be resized, data augmentation strategies are used and CLAHE is employed to improve the image quality.
- Following that, features such as moment invariant features, GLCM features, and wavelet features are extracted from the preprocessed image.
- The EfficientNet classifier is trained using these features to categorize the cervical images as Normal or Abnormal.
- Finally, we propose a SegNet for segmenting the defected areas, it uses multi-stage architecture and attention blocks in each stage.
- The Herlev dataset has undergone various ablation experiments. Our proposed network outperforms the state efficiency concerning all other approaches, according to the experimental data.

The paper is organized as follows. In Section II, a few similar prior efforts are summarized. Section III describes the proposed system. Section IV presents experimental findings and a discussion. Section V contains the work's conclusion.

## II. LITERATURE REVIEW

In the literature, there is a lot of research comparing the effectiveness of various methods utilized to treat cervical cancer. ML and DL techniques were used in group studies. It is clear from studies on cervical cancer that deep learning techniques including CNN, stacked autoencoder, VGG19, and LASSO were applied.

Convolutional neural networks (CNNs) were introduced by Ghoneim et al. [21] for the identification and categorization of cervical cancer cells. To extract deep-learned features, a CNNs model is fed the cell pictures. After that, a classifier powered by an extreme learning machine (ELM) classifies the input photographs. Through transfer learning and fine-tuning, CNN's model is applied. Their accuracy is superior to others as compared to the current system. Comparatively, the level of complexity is higher.

To classify cervical cancer from Pap smears, William et al. [22] developed an improved fuzzy c-means method. Through the employment of a Trainable Weka Segmentation classifier, cells were segmented, and trash was eliminated sequentially. Wrapper filters were used to select features. The method surpasses several of the current algorithms in terms of false negative rate, false positive rate, and classification error, according to the results. The primary drawback of the recommended automated Pap smear analysis systems is their inability to handle the Pap smear architectures complexities.

Deep learning methods, such as softmax classification with stacked autoencoder, have reportedly been utilized to

categorize data sets, according to Adem et al. [23]. The raw data collection is transformed into a lower dimension dataset by applying a stacked autoencoder. In order to minimize the data dimension and create a classifier with high accuracy, the stacked autoencoder model was used. Comparatively speaking to the other machine learning method, it has higher classification success rates for data related to cervical cancer. It is necessary to increase accuracy by removing relevant information.

The VGG19 (TL) model and the CYENET were presented by Chandran et al. [24] to automatically classify cervical tumors from colposcopy pictures. By improving the VGG19 model, which is extensively used for medical image processing, the transfer learning method is applied to forecast accuracy. By utilizing an optimal architecture and an ensemble technique CYENET, the CNN created from scratch is intended to automate the screening of cervical pictures. The outcomes of the experiments demonstrate that the suggested CYENET had high performances. As a result of the dimension reduction, training process is very long.

The classification of cervical biopsy tissue images based on LASSO and ensemble learning-support vector machine (EL-SVM) was first presented by Huang et al. [25]. The average optimization time was decreased by 35.87 seconds while maintaining the classification accuracy when the LASSO technique was used for feature selection. Serial fusion was then carried out. 468 biopsy tissue pictures were identified and classified using the EL-SVM classifier. The ROC curve and error curve were utilized to assess the classifier's generalizability. The results of the experiment indicate that a superior categorization result was obtained. A two-step feature selection process that takes time and makes it challenging to distinguish between individual cells.

A smaller visual Geometry Group-like Network is used to classify the segmented entire cervical cell data by Allehaibi et al. [26] using a Mask R-CNN and VGG-like Net. The ResNet10 network serves as the foundation of the Mask R-CNN, fully utilizing geographical data and past knowledge. Mask R-CNN performs better in precision, recall, and ZSI than the prior segmentation approach during the segmentation phase when applied to the entire cell. The performance of the seven-class problem categorization produces excellent results. The suggested method uses a little volume of data and requires more research to fully understand cervical cells.

Allehaibi et al. [27] presented an Inception v3-based cervical cell categorization system with features that were intentionally extracted. The accuracy of cervical cell recognition has been significantly improved by the use of Inception v3 and artificial characteristics. Additionally, this research inherits the strong learning capability from transfer learning to produce an accurate and efficient classification of cervical cell images while addressing the under-fitting issue with a limited amount of medical data. The suggested algorithm offers great accuracy, good universality, and minimal complexity. The Suggested method had issues with certain cells having overlapping cytoplasm sections. Table I represents the merits and demerits of the existing papers.

TABLE I. MERITS AND DEMERITS OF RELATED WORKS

Reference	Year	Method	Dataset	Merits	Demerits
[21]	2020	CNN& Extreme Machine learning	Herlev database	More scalable and practical	Investigating cervical cells requires more research.
[22]	2019	fuzzy c-means algorithm	DTU/Herlev benchmark Pap smear dataset	Higher precision and smaller data dimensions	Due to the dimension reduction, training time is quite long.
[23]	2019	stacked autoencoder	Cervical cancer dataset	Acquire more complementing features	an increase in image fusion complexity
[24]	2021	VGG19 and CYENET	Intel ODT dataset	Better accuracy Efficient classification	More complexity Need more investigation
[25]	2020	LASSO- (EL-SVM)	Cervical cancer	Better sensitivity and specificity	Extraction of pertinent data is required to increase accuracy.
[26]	2019	Mask R-CNN	Herlev Pap Smear dataset	Improvements in accuracy and feature selection	The selection of features in two phases takes time.
[27]	2019	Inception v3	Herlev dataset	Excellent robustness and best performance	need development to change the parameter

It is possible to draw a conclusion using a few of the restrictions and potential improvements that may be learned from the existing research. The effectiveness of existing approaches was insufficient. The algorithm's complexity and potential are primarily responsible for this. Our proposed approach is used to address the gaps in the current body of research. The segmentation and classification method for cervical cancer that has been suggested is intended to enhance classifier performance.

### III. METHODOLOGY

The four essential components of the proposed methodology were preprocessing, feature extraction, classification, and segmentation. The preprocessing processes are used to increase the classification efficiency even more. The raw photos were prepared using the pre-processing method. Data compression and image enhancement were part of the pre-processing. By using the CLAHE approach the

image can be enhanced. Then, characteristics including GLCM features, wavelet features, and moment invariant features are extracted from the preprocessed image. By comparing the cervical picture with the taught features, the EfficientNet classifier is trained to determine if the image is normal or abnormal. Finally, SegNet is employed to segment the defected area. After that, the dataset is separated into three parts. The training model is then fed this dataset. A selection of test datasets is utilized to label the training model, which is then used to categorize cervical cancer. Fig. 1 demonstrates the architecture diagram of the proposed framework.

#### A. Preprocessing

Images are strengthened in resolution during the initial stages of pre-processing utilizing various filters. To prevent further distortions, several data augmentation methods, including shearing, scaling, and rotation, are also used. Images are also resized at various scales. The CLAHE algorithm can be used to improve the image.

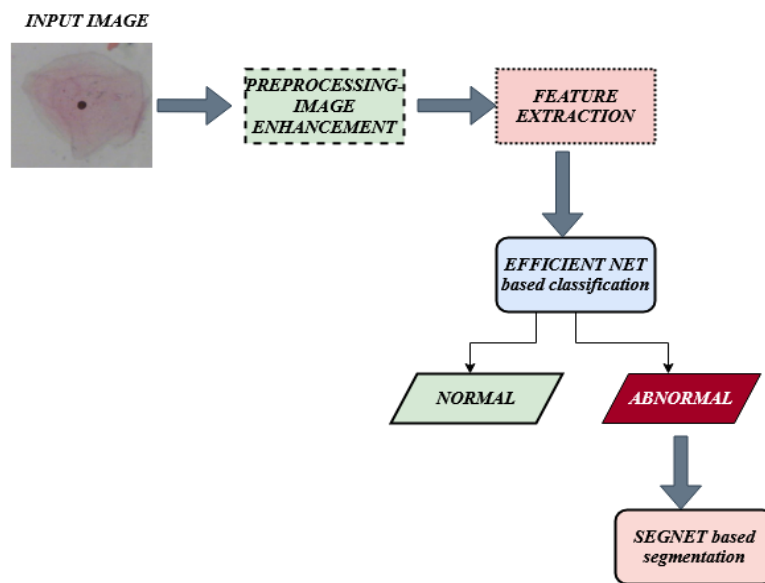


Fig. 1. Architecture Diagram of the Proposed Methodology.

1) *Contrast Local Adaptive Histogram Equalization (CLAHE)*: With the help of histogram clipping, the histogram-based image enhancement technique known as CLAHE can only amplify images to a certain degree. It is a technique that is efficient for assigning projected intensity levels in medical data. In order to adjust the brightness of a pixel's display to reflect where that pixel ranks in terms of intensity in its histogram, the method analyses an intensities histogram in a contextual region that is focused on each individual pixel. The histogram that is generated shows the image contrast created by the technique at every luminance. It is a modified version of the standard histogram.

It lessens contrast enhancement, which is typically achieved through the histogram equalization method, which also makes more noise. As a consequence, by reducing contrast augmentation in Histogram Equalization, the expected result was attained in situations where noise became particularly obvious by enhancing contrast, such as medical photos. To lessen contrast, one can restrict the slope of the linked function. The use of CLAHE in our work has helped to increase the accuracy rate overall.

#### B. Feature Extraction

The nucleus' and cytoplasm's essential characteristics, including texture, shape, and color, were retrieved at this stage. In an image, features stand for the traits of the pixel pattern. In this study, morphological, wavelet, and GLCM characteristics are retrieved from cervical images to distinguish between normal and pathological images. GLCM was employed to extract eight texture features. Normal and abnormal cells appear very differently in the cervical Pap smear image in terms of color and form distribution.

2) *Wavelet features*: Due to their speed and superior transformation capabilities compared to other transforms like Contourlet and Curvelet, Wavelets are beneficial in multi-resolution analysis of cervical images. This study decomposes the magnitude response Gabor image using the Discrete Wavelet Transform (DWT), which is applied to every row and column.

The LL, HL, LH, and HH sub-bands are produced by the first level decomposition. L and H stand for low and high frequencies, respectively. Additionally, the second-level decomposition of DWT is applied to the LL sub-band to create four additional subbands. The feature pattern for the cervical image classifications uses each of these subbands.

3) *GLCM features*: The feature extraction method known as GLCM is employed to extract the energy features of the cervical picture. Any single channel image can have one built. The GLCM is a square matrix with the same number of rows and columns as there were in the original image's grayscale. The GLCM matrix is created by counting the number of times a grayscale intensity pixel will be found next to a pixel with the value of the fused cervical picture at various orientations, such as 0°, 45°, 90°, and 135°. The GLCM matrix is built in this work at a 45° angle. The contrast, energy, entropy, and

correlation characteristics of the GLCM features are employed to distinguish the cancerous image from the healthy cervical image.

Contrast: It's described as,

$$Contrast = \sum(|i - j|^2 \times p(i, j)) \quad (1)$$

This texture feature calculates the difference in grayscale between adjacent pixels.

Correlation: It's described as,

$$Correlation(R) = \frac{\sum(i - \mu_i)(j - \mu_j) \cdot p(i, j)}{[\sigma_i, \sigma_j]} \quad (2)$$

The correlation between brightness in adjacent pixels is measured by R. The GLCM's row average  $\mu_i$  and column average ( $\mu_j$ ) are respectively. The GLCM's row  $\mu_i$  and column  $\mu_j$  corresponding standard deviations are denoted by  $\sigma_i$  and  $\sigma_j$ .

Energy: It is a second angular moment, calculated by adding the squares of all the GLCM's components.

$$Energy(E) = \sum p(i, j)^2 \quad (3)$$

Energy, a unit of measurement for homogeneity, runs from zero to unity for a picture.

Entropy: It is described as,

$$Entropy = -\sum p(i, j)[\log_2 p(i, j)] \quad (4)$$

The degree to which the GLCM's elements are near the diagonal is gauged by this metric. The value is between 0 and 1.

4) *Morphological features*: Cell size and form make up the morphological characteristics. For this study, eight connected chain codes were used to determine the morphological characteristics of cells. The eight pixels around the eight-linked chain code are the connected pixels.

a) *Area*: the number of pixels that each cell has taken up;

b) *Circumference*: The cell's circumference is equal to one week.

c) *Nuclear to cytoplasm ratio*: ratio of nuclear to cellular size:

$$\frac{N}{C} = \frac{Nucl_{area}}{Nucl_{area} + Cyto_{area}} \quad (5)$$

#### C. Classification based on EfficientNet

The EfficientNet group has eight variants, spanning from B0 to B7, and the quantity of estimated parameters does not increase much as the amount of models increases, even though accuracy increases. In contrast to previous CNN models, EfficientNet employs the Swish activation function rather than the Rectifier Linear Unit (ReLU) activation function. Deep learning frameworks seek to uncover more efficient methods using fewer methods. EfficientNet, unlike other state-of-the-

art approaches, achieves more efficient results by scaling depth, width, and resolution evenly while reducing the strategy's size. When resources are limited, the initial step in the compound scaling strategy is to discover the relationship among the various scaling dimensions of the network. This method determines an appropriate scaling factor for the depth, breadth, and resolution dimensions. After that, these coefficients are used to scale the baseline network to the target network. Table II represents the EfficientNet architecture.

CNN belongs to the EfficientNet group. In terms of layer width, layer depth, input resolution, and a combination of these criteria, EfficientNet methods scale well. EfficientNet is a recent deep learning approach that aims to improve model efficiency while also improving accuracy. From B0 to B7, there are various variants. The inversion bottleneck MB Conv is the basic building piece for EfficientNet. It was first presented in MobileNetV2, however, it is used significantly further than MobileNetV2 because of the larger FLOPS budget. Blocks in MBConv are made up of layers that expanded and then compress the channels, hence direct connections are employed among bottlenecks that connected far fewer channels than expansion layers. When compared to typical layers, this design has in-depth separate convolutions that minimize computation by almost a k2 factor. The 2D convolution window's height and breadth are determined by the kernel size, or k, which is the opposite.

EfficientNet presents a novel compound scaling model that scales network width, depth, and image size uniformly using a compound coefficient  $\phi$ .

$$\text{depth: } d = \alpha \phi \tag{6}$$

$$\text{width: } w = \beta \phi \tag{7}$$

$$\text{resolution: } r = \gamma \phi \tag{8}$$

$$\begin{aligned} \text{s.t. } & \alpha \cdot \beta^2 \cdot \gamma^2 \approx 2 \\ & \alpha \geq 1, \beta \geq 1, \gamma \geq 1 \end{aligned} \tag{9}$$

FLOPS are proportional to d, w2, and r2 in a standard convolution process. Because convolution operations account for the majority of the cost of computation in convolution networks, growing the network as indicated in Eq. (3.5) boosts the network's FLOPS by about  $(\alpha, \beta^2, \gamma^2) \phi$  in total.

The batch normalization (BN) restricts the final layer's outcome to a range, requiring a mean of zero and one SD. This adjustment shortens the training period and improves the model's stability. The compound scaling approach scales this system in two phases, starting with the baseline EfficientNet-B0.

Step 1: Considering double as many materials are allocated, a grid search with  $\phi = 1$  is used to find the best values for  $\alpha, \beta, \gamma$ .

Step 2: The generated  $\alpha, \beta, \gamma$  values are set as constants, and the baseline network is scaled up using Equation. (3.5) with varied values  $\phi$  to generate EfficientNet-B1 through B7.

#### D. Segmentation

Using morphological techniques, the cancerous regions of an aberrant cervical picture are segmented. An encoder network, a decoder network, and a final layer for pixel-wise categorization make up SegNet [28]. This configuration consists of four blocks, the final block of which does not execute pooling. To create feature vectors that correspond to each input, features from the input image are extracted using the encoder layers. A decoder network that consists of four blocks of upsampling, convolution, and batch normalizing layers is then applied after that.

TABLE II. EFFICIENTNET FRAMEWORK

Level	Operator	Resolution	Channels	Layers
<b>EfficientNetB0 architecture, the network baseline</b>				
1	Conv1×1/Pool/FC	7×7	1,280	1
2	MBCConv6, k3×3	7×7	320	1
3	MBCConv6, k6×6	14×14	192	4
4	MBCConv6, k3×2	14×14	112	3
5	MBCConv6, k5×4	28×28	80	3
6	MBCConv6, k5×5	56×56	40	2
7	MBCConv6, k3×3	112×112	24	2
8	MBCConv1, k3×3	112×112	16	1
9	Conv 3×3	224×224	32	1
<b>Additional layers</b>				
10	FC/Softmax	1	NC	1
11	FC/BN/Swish	1	128	1
12	FC/BN/Swish/Dropout	1	512	1
13	B.N./Dropout	7×7	1280	1

Segmentation mask is created by the decoder network using feature vectors, and output at high resolution is produced by upsampling layers using low features. The final layer is a classification layer that uses 2D convolution with a 1x1 filter size to do pixel-by-pixel classification. Throughout the training, a stride of one and a filter size of 3x3 are employed. Except for the final layer, where the sigmoid activation function is employed, ReLU is employed as the activation function. To reduce the size of the feature map, max-pooling layers with a pool size of 2x2 are employed [29]. After each batch, the weights are optimized by the Adam optimizer with a learning rate of 1e-4.

The model is trained throughout 10 epochs with a batch size of 32. For pixel-wise segmentation, a Softmax classifier is given the final decoder output feature maps. For quick and precise image segmentation, the decoder recovered spatial dimensions. Due to its memory and processing speed, the SegNet architecture is generally superior to other systems like U-Net and FCN.

#### IV. RESULT AND DISCUSSIONS

To illustrate the conclusion, using a benchmark dataset to compare the proposed method to existing techniques in terms of NPV, sensitivity, specificity, PPV, and accuracy. The materials and metrics that were employed to achieve the intended results will be described in this paper. The proposed experiment's performance was evaluated in PYTHON.

##### A. Dataset Description

Herlev dataset, which is made available to the public for disease detection, was obtained from Denmark Hospital to detect cervical disease. There are 917 total Pap smear images in the complete data set. The dataset is split into a training set and a testing set, with the training set including 643 photos and the testing set including 274 images. The classifications of Normal and Abnormal have been taken into consideration.

With a total cell count of 917, we have taken into account the various cervical cell types, including epithelial and dysplastic, which are divided into normal and abnormal classes. The sample smear cervical cells from the dataset shown in Fig. 2 were chosen at random.

##### B. Evaluation Metrics in Cervical Cancer Diagnosis

The effectiveness of the categorization model is explained by several evaluation criteria used by various authors. Confusion matrices are used to assess the model's efficacy for the majority of medical image classification. Sensitivity, Specificity, Accuracy, and F1 score are some of the distinctive metrics employed for the analysis. The prediction output includes four results: true positive, true negative, false positive, and false negative. The various performance criteria used to evaluate the mode are displayed in Table III.

Accuracy: The number of correctly identified images determines a technique's classification accuracy, which is evaluated as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (10)$$

Sensitivity (Recall): It measures the percentage of positive samples that are accurately categorized. Sensitivity has a value between 0 and 1.

$$Sensitivity = \frac{TP}{TP + FN} \quad (11)$$

Specificity: It is a measurement of the percentage of incorrectly identified negative samples.

$$Specificity = \frac{TN}{TN + FP} \quad (12)$$

Positive Predictive Value (PPV): It counts the number of pixels that are positive and accurately identify cancerous regions.

$$PPV = \frac{TP}{TP + FP} \quad (13)$$

Negative Predictive Value (NPV): It counts the number of negative pixels that are associated with incorrectly identified cancer region pixels.

$$NPV = \frac{TN}{TN + FN} \quad (14)$$

Confusion Matrix: The Confusion Matrix summarizes the categorization problem's prediction results. The confusion matrix reveals not just the classifier's errors, but also the sorts of errors. Fig. 3 represents the output of cervical cancer.

##### C. Evaluation of Classification Performances

A comparison of classification techniques and current classifiers is provided in this section. Four classification methods AlexNet, LeNet, VGG and Inception V3 are used in this study. Table III displays the effectiveness of many strategies, including the one that is suggested.

The existing approaches like Lenet, AlexNet, VGG and Inception V3 are compared with the proposed approach. When comparing with the sensitivity metrics it achieves 89.73% in LeNet, 90.16% in AlexNet, 91.37% in VGG, 99.44% in Inception V3 and our proposed approach yield a greater solution which is 99.67%. The next comparison can be made in Specificity LeNet achieves 84.33%, AlexNet yields 87.34%, VGG gains 86.88%, 96.73% in Inception V3 and the proposed approach yields 98.39%.

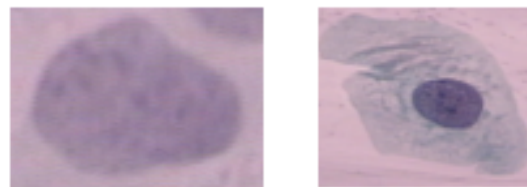


Fig. 2. Sample Images from the Dataset.

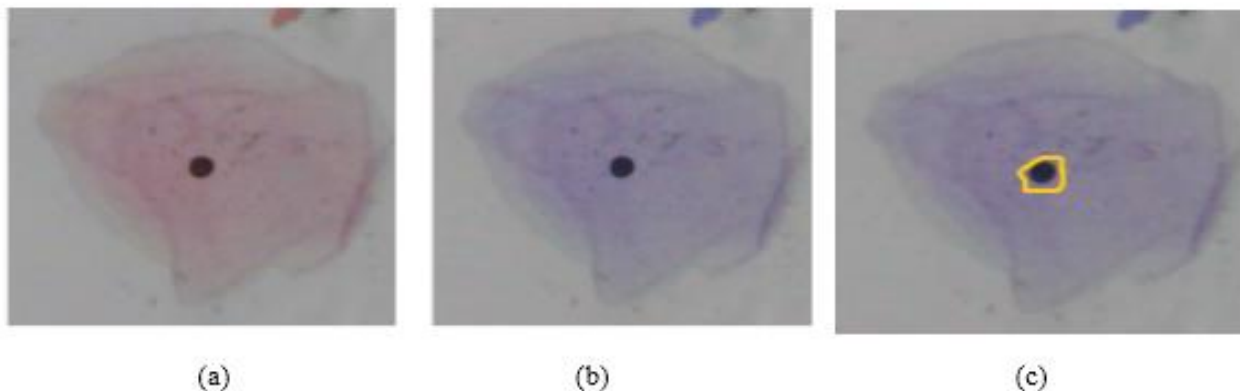


Fig. 3. Output of Cervical Cancer Affected Image (a) Input Image (b) Contrast-enhanced (c) Segmented Defected Area.

TABLE III. PERFORMANCES COMPARISON OF PROPOSED WITH EXISTING APPROACHES

Approaches	Sensitivity	Specificity	Accuracy
LeNet	89.73	84.33	86.76
AlexNet	90.16	87.34	89.57
VGG	91.37	86.88	91.47
Inception V3	99.44	96.73	98.23
EfficientNet	99.67	98.39	99.05

TABLE IV. COMPARISON OF PERFORMANCES OF THE CLASSIFIERS

Approaches	Accuracy (%)	Specificity (%)	Sensitivity (%)	PPV (%)	NPV (%)
CYENET	92.30	96.20	92.40	92	95
DenseNet-121	72.42	76.83	59.86	48.39	84.52
DenseNet-169	69.79	71.48	65	44.84	85.31
SVM	63.27	71.85	78.46	70	76.87
Proposed	99.67	98.39	99.67	94.67	96.13

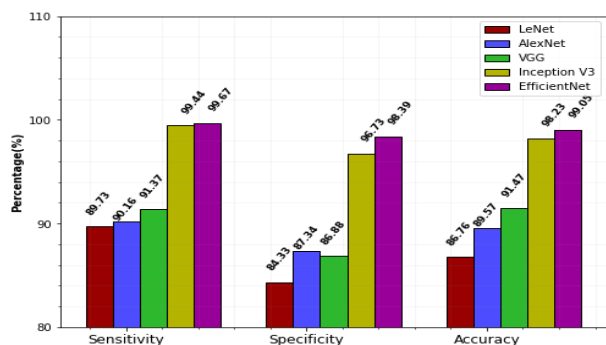


Fig. 4. Performances Comparison of Proposed with Existing Approaches.

Finally, a comparison can be made with accuracy metrics our proposed approach yields 99.05% which is the best solution then the worst solution appear in LeNet which is 86.76%. Performances comparison of proposed with existing approaches is represented in Fig. 4.

The metrics like Accuracy, Specificity, Sensitivity, PPV and NPV are used to compare the performances. To compare our proposed approach performances, the existing approach includes CYENET, DenseNet-121, DenseNet-169 and SVM utilized (Table IV).

Comparison can be made with the Accuracy metrics CYENET achieves 92.30% of accuracy, DenseNet-121 yields 72.42% of accuracy, DenseNet-169 gains 69.79% of accuracy, SVM achieves 63.27% of accuracy finally our proposed approach gains 99.67% of accuracy which is the greater one. Fig. 5 represents the Accuracy comparison of the proposed with existing approaches.

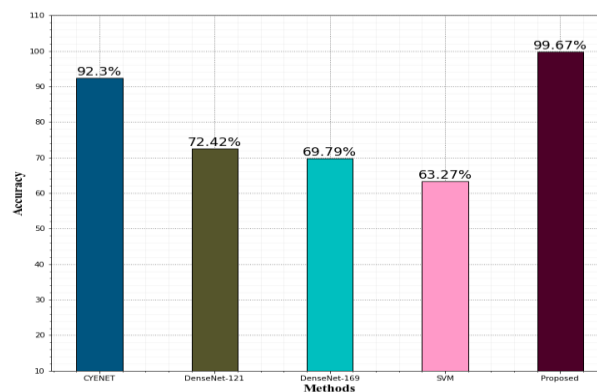


Fig. 5. The Accuracy Comparison of Proposed with Existing Approaches.

When comparing with specificity our proposed approach gains 98.39%, CYENET achieves 96.20%, DenseNet-121 yields 76.83%, DenseNet-169 gains 71.48%, and SVM achieves 71.85%. The figure represents the specificity comparison of the proposed with the existing. The Specificity performance comparison of the proposed with existing approaches is shown in Fig. 6.

Comparison can be made with the Sensitivity metrics CYENET achieves 92.4% of Sensitivity, DenseNet-121 yields 59.86% of Sensitivity, DenseNet-169 gains 65% of Sensitivity, SVM achieves 78.46% of Sensitivity finally our proposed approach gains 99.67% of Sensitivity which is the greater one. Fig. 7 represents the sensitivity comparison of the proposed with existing.



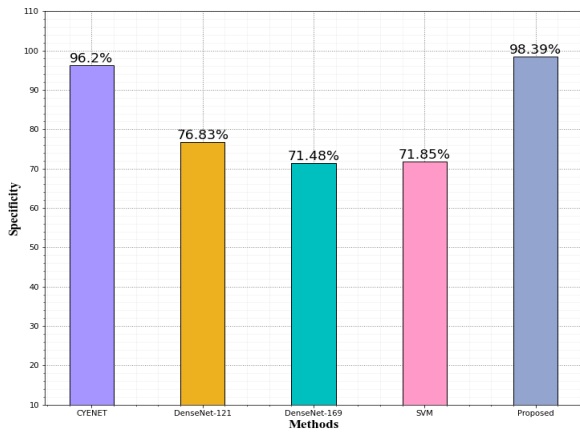


Fig. 6. The Specificity Performances Comparison of Proposed with Existing Approaches.

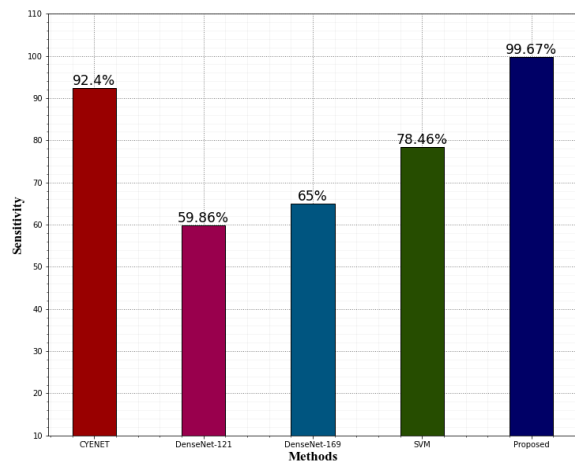


Fig. 7. The Sensitivity Performances Comparison of the Proposed with Existing Approaches.

Comparison can be made with the Sensitivity metrics CYENET achieves 92% of PPV, DenseNet-121 yields 48.39% of PPV, DenseNet-169 gains 44.84% of PPV, SVM achieves 70% of PPV finally our proposed approach gains 94.67% of PPV which is the greater one. Fig. 8 represents the PPV comparison of proposed with existing.

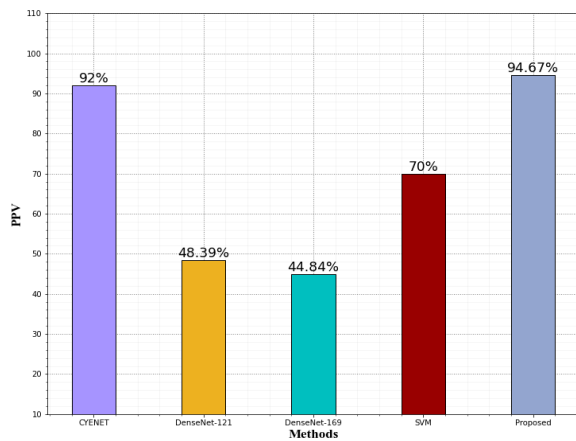


Fig. 8. The PPV Metrics Comparison of Proposed with Existing Approaches.

Comparison can be made with the Sensitivity metrics CYENET achieves 95% of NPV, DenseNet-121 yields 84.52% of NPV, DenseNet-169 gains 85.31% of NPV, SVM achieves 76.87% of NPV finally our proposed approach gains 94.67% of NPV which is the greater one. The Fig. 9 represents the NPV comparison of the proposed with the existing.

The confusion matrix for the end-to-end trained proposed method is shown in Fig. 10. 3% of Normal, 2% of abnormal samples were misclassified, while 97 % of benign samples, 98% of cancerous samples were classified correctly. As a result, the proposed approach acquires the best result.

#### D. Evaluation of Training and Testing

Train Accuracy and Validation Accuracy curves converge in the end, and after 50 epochs we received an accuracy of 99.56%, which is quite good. The validation Loss curve jumps up and down a bit. It means it would be nice to have more validation data. Fig. 11 represents the proposed training accuracy versus testing accuracy.

After about 25 epochs Validation Loss exceeds Train Loss, which means we have a bit of overfitting here. But the curve doesn't go up over epochs, and the difference between Validation and Train Loss is not that big, so this could be accepted. Fig. 12 represents the proposed training loss versus testing loss

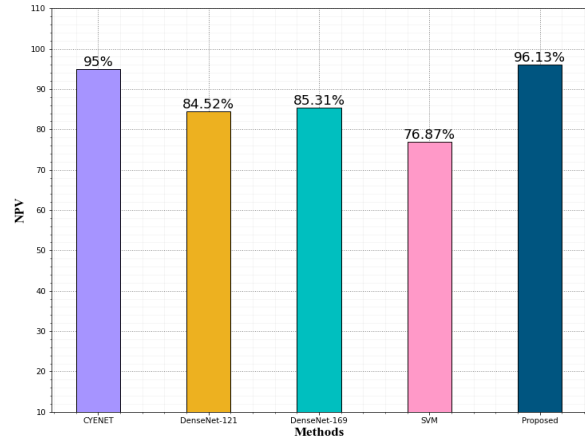


Fig. 9. The NPV Metrics Comparison of Proposed with Existing Approaches.

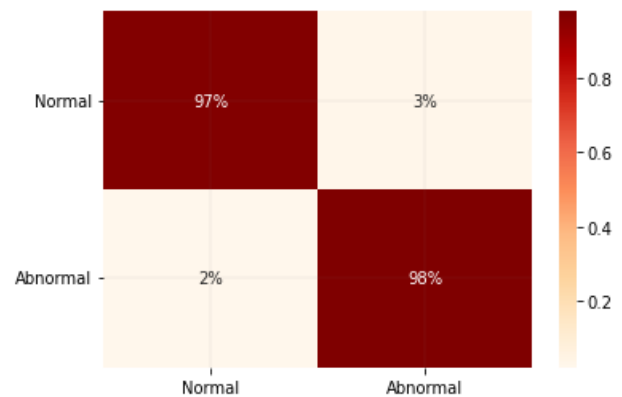


Fig. 10. Confusion Matrix.

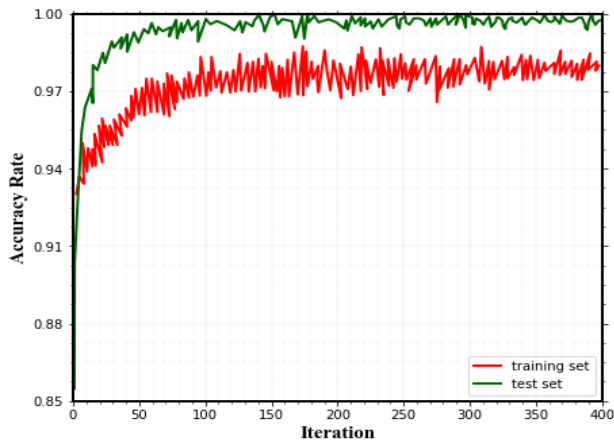


Fig. 11. Proposed Training Accuracy Versus Testing Accuracy.

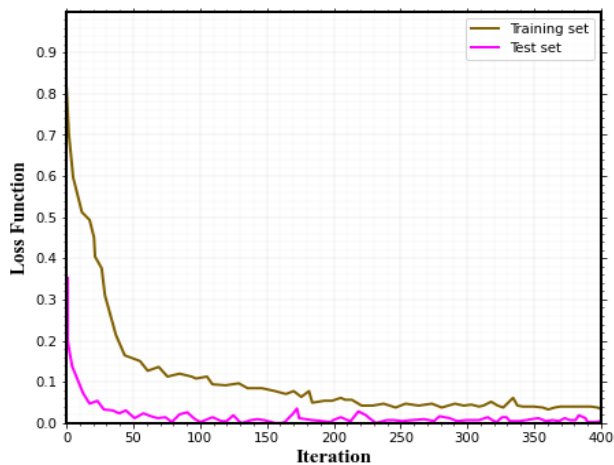


Fig. 12. Proposed Training Loss Versus Testing Loss.

### E. Discussions

The success rate of treatment for cervical cancer is greatly impacted by early identification. The pathological study of microscopic Pap smear slide pictures is the primary cervical cancer screening tool. Automatic image analysis techniques are likely to defeat subjective justifications and lighten the workload. Due to the considerable unpredictability of cervical cell pictures, including overlapping cells, dust, contaminants, and uneven irradiation, effective nucleus image segmentation remains a difficult challenge. Furthermore, restrictions in feature design and selection make it difficult to classify cervical smear images.

The quantitative analysis of microscopic Pap smear slide pictures is difficult as a result. The absence of uninvolved photos from the public database of cervical images is crucial for the early detection of cervical cancer. Additionally, there are few cervical smear photos that have been labelled. The paper proposed an automatic cervical smear image categorization system based on EfficientNet to address the existing issues mentioned above. The usefulness of the proposed approach, which takes time to apply, is shown by experimental findings. Future research might focus on improving the method's efficiency and lowering computational complexity.

### V. CONCLUSION

In conclusion, a set of clinically relevant and biologically understandable features are used to offer an automated detection and classification approach for the identification of cancer from cervical pictures. The proposed methodology is based on a network for segmenting and classifying cancer. The CLAHE-based approach is utilized to improve the cervical pictures. EfficientNet classifier is employed to divide cervical pictures into normal and abnormal images. According to the simulation results, the suggested cervical cancer segmentation method can identify both normal and abnormal areas in images of the cervical region. The cervical cancer detection system's performance metrics are 97.42 percent sensitivity, 99.36 percent specificity, 98.29 percent accuracy, 97.28 percent PPV, and 92.17 percent NPV. The theoretical deep learning model will be tested on other datasets in the future. Combining a few sophisticated image processing techniques with the method can also improve it.

### ACKNOWLEDGMENT

We declare that this manuscript is original, has not been published before and is not currently being considered for publication elsewhere.

### REFERENCES

- [1] Wentzensen, N., Lahrmann, B., Clarke, M. A., Kinney, W., Tokugawa, D., Poitras, N., ... & Grabe, N. (2021). Accuracy and efficiency of deep-learning-based automation of dual stain cytology in cervical Cancer screening. *JNCI: Journal of the National Cancer Institute*, 113(1), 72-79.
- [2] Mohammadi, R., Shokatian, I., Salehi, M., Arabi, H., Shiri, I., & Zaidi, H. (2021). Deep learning-based auto-segmentation of organs at risk in high-dose rate brachytherapy of cervical cancer. *Radiotherapy and Oncology*, 159, 231-240.
- [3] Alyafei, Z., & Ghouti, L. (2020). A fully-automated deep learning pipeline for cervical cancer classification. *Expert Systems with Applications*, 141, 112951.
- [4] Matsuo, K., Purushotham, S., Jiang, B., Mandelbaum, R. S., Takiuchi, T., Liu, Y., & Roman, L. D. (2019). Survival outcome prediction in cervical cancer: Cox models vs deep-learning model. *American journal of obstetrics and gynecology*, 220(4), 381-e1.
- [5] Chandran, V., Sumithra, M. G., Karthick, A., George, T., Deivakani, M., Elakkiya, B., ... & Manoharan, S. (2021). Diagnosis of cervical cancer based on ensemble deep learning network using colposcopy images. *BioMed Research International*, 2021.
- [6] Jiang, X., Li, J., Kan, Y., Yu, T., Chang, S., Sha, X., ... & Wang, S. (2020). MRI-based radiomics approach with deep learning for prediction of vessel invasion in early-stage cervical cancer. *IEEE/ACM transactions on computational biology and bioinformatics*, 18(3), 995-1002.
- [7] Kudva, V., Prasad, K., & Guruvare, S. (2017). Detection of specular reflection and segmentation of cervix region in uterine cervix images for cervical cancer screening. *Irbm*, 38(5), 281-291.
- [8] Wu, M., Yan, C., Liu, H., Liu, Q., & Yin, Y. (2018). Automatic classification of cervical cancer from cytological images by using convolutional neural network. *Bioscience reports*, 38(6).
- [9] Ch, P. N., Gurram, L., Chopra, S., & Mahantshetty, U. (2018). The management of locally advanced cervical cancer. *Current opinion in oncology*, 30(5), 323-329.
- [10] Lee, J., Chang, C. L., Lin, J. B., Wu, M. H., Sun, F. J., Jan, Y. T., ... & Chen, Y. J. (2018). Skeletal Muscle Loss Is an Imaging Biomarker of Outcome after Definitive Chemoradiotherapy for Locally Advanced Cervical Cancer. *Skeletal Muscle Loss in Cervical Cancer*. *Clinical Cancer Research*, 24(20), 5028-5036.
- [11] Kudva, V., Prasad, K., & Guruvare, S. (2020). Transfer learning for classification of uterine cervix images for cervical cancer screening. In

- Advances in Communication, Signal Processing, VLSI, and Embedded Systems (pp. 299-312). Springer, Singapore.
- [12] Melamed, A., Margul, D. J., Chen, L., Keating, N. L., Del Carmen, M. G., Yang, J., ... & Rauh-Hain, J. A. (2018). Survival after minimally invasive radical hysterectomy for early-stage cervical cancer. *New England Journal of Medicine*, 379(20), 1905-1914.
- [13] Asadi, F., Salehnasab, C., & Ajori, L. (2020). Supervised algorithms of machine learning for the prediction of cervical cancer. *Journal of biomedical physics & engineering*, 10(4), 513.
- [14] Singh, S. K., & Goyal, A. (2020). Performance analysis of machine learning algorithms for cervical cancer detection. *International Journal of Healthcare Information Systems and Informatics (IJHISI)*, 15(2), 1-21.
- [15] Kudva, V., Prasad, K., & Guruvare, S. (2020). Hybrid transfer learning for classification of uterine cervix images for cervical cancer screening. *Journal of digital imaging*, 33(3), 619-631.
- [16] Kan, Y., Dong, D., Zhang, Y., Jiang, W., Zhao, N., Han, L., ... & Luo, Y. (2019). Radiomic signature as a predictive factor for lymph node metastasis in early-stage cervical cancer. *Journal of Magnetic Resonance Imaging*, 49(1), 304-310.
- [17] Matsuo, K., Machida, H., Shoupe, D., Melamed, A., Muterspach, L. I., Roman, L. D., & Wright, J. D. (2017). Ovarian conservation and overall survival in young women with early-stage cervical cancer. *Obstetrics and gynecology*, 129(1), 139.
- [18] Jia, A. D., Li, B. Z., & Zhang, C. C. (2020). Detection of cervical cancer cells based on strong feature CNN-SVM network. *Neurocomputing*, 411, 112-127.
- [19] Nirmal Jith, O. U., Harinarayanan, K. K., Gautam, S., Bhavsar, A., & Sao, A. K. (2018). DeepCerv: Deep neural network for segmentation free robust cervical cell classification. In *Computational Pathology and Ophthalmic Medical Image Analysis* (pp. 86-94). Springer, Cham.
- [20] Ghoneim, A., Muhammad, G., & Hossain, M. S. (2020). Cervical cancer classification using convolutional neural networks and extreme learning machines. *Future Generation Computer Systems*, 102, 643-649.
- [21] Ghoneim, A., Muhammad, G., & Hossain, M. S. (2020). Cervical cancer classification using convolutional neural networks and extreme learning machines. *Future Generation Computer Systems*, 102, 643-649.
- [22] William, W., Ware, A., Basaza-Ejiri, A. H., & Obungoloch, J. (2019). Cervical cancer classification from Pap-smears using an enhanced fuzzy C-means algorithm. *Informatics in Medicine Unlocked*, 14, 23-33.
- [23] Adem, K., Kiliçarslan, S., & Cömert, O. (2019). Classification and diagnosis of cervical cancer with stacked autoencoder and softmax classification. *Expert Systems with Applications*, 115, 557-564.
- [24] Chandran, V., Sumithra, M. G., Karthick, A., George, T., Deivakani, M., Elakkiya, B., ... & Manoharan, S. (2021). Diagnosis of cervical cancer based on ensemble deep learning network using colposcopy images. *BioMed Research International*, 2021.
- [25] Huang, P., Zhang, S., Li, M., Wang, J., Ma, C., Wang, B., & Lv, X. (2020). Classification of cervical biopsy images based on LASSO and EL-SVM. *IEEE Access*, 8, 24219-24228.
- [26] Jia, A. D., Li, B. Z., & Zhang, C. C. (2020). Detection of cervical cancer cells based on strong feature CNN-SVM network. *Neurocomputing*, 411, 112-127.
- [27] Allehaibi, K. H. S., Nugroho, L. E., Lazuardi, L., Prabuwono, A. S., & Mantoro, T. (2019). Segmentation and classification of cervical cells using deep learning. *IEEE Access*, 7, 116925-116941.
- [28] Almotairi, S., Kareem, G., Aouf, M., Almotairi, B., & Salem, M. A. M. (2020). Liver tumor segmentation in CT scans using modified SegNet. *Sensors*, 20(5), 1516.
- [29] Weng, L., Xu, Y., Xia, M., Zhang, Y., Liu, J., & Xu, Y. (2020). Water areas segmentation from remote sensing images using a separable residual segnet network. *ISPRS International Journal of Geo-Information*, 9(4), 256.

# Cloud based Forecast of Municipal Solid Waste Growth using AutoRegressive Integrated Moving Average Model: A Case Study for Bengaluru

Rashmi G<sup>1</sup>

Department of Computer Science & Engineering  
Research Scholar, RNSIT  
Bengaluru, India

S Sathish Kumar K<sup>2</sup>

Department of Information Science & Engineering  
Professor, RNSIT  
Bengaluru, India

**Abstract**—Forecasting the quantity of waste growth in upcoming years is very much required for assessing the existing waste management system. In this research work, time series forecast model, ARIMA (Autoregressive Integrated Moving Average), is used to predict future waste growth from 2021 to 2028 for Bengaluru, largest city in Karnataka. Eight years old historical solid waste dataset from 2012 to 2020 is used to make predictions. This dataset is preprocessed and only time bounded variables like days, month, year and waste quantity in tons are used in this research work to obtain accurate prediction. The model is implemented in python in Google Colab free cloud’s Jupyter notebook. As ARIMA is time bounded, forecast made by the model is accurate and performance of the model is evaluated using metrics such as Mean Absolute Deviation (MAD), Mean Absolute Percentage Error (MAPE), Root Mean Square Error (RMSE) and Coefficient of Determination ( $R^2$ ). Outcomes revealed that ARIMA (0, 1, 2) model with the lowermost RMSE (753.5742), MAD (577.4601), and MAPE (11.6484) values and the maximum  $R^2$  (0.9788) value has a greater forecast performance. The outcomes attained from the model also showed that the total volume of yearly solid waste to be produced will rise from about 50,300 tons in 2021 to 75,600 tons in 2028.

**Keywords**—Cloud Computing; Machine Learning; Time Series Forecasting; Waste Management System; ARIMA; Predictive Modeling

## I. INTRODUCTION

Today’s global technologies are popularly driven by cloud computing and machine learning. Both of these are contributing to every organization’s business growth. Machine learning today facilitates users to create models which can be used to make predictions by training them to automatically learn from past data. Various machine learning approaches [7] such as supervised and unsupervised require huge amount of storage which is a challenging task for machine learning professionals. Cloud computing [9] contributes in such scenarios by providing all the resources and services required to ease the tasks. Machine learning makes brainy applications where as cloud computing provides storing and refuge services to access these applications. Cloud computing [2] thus helps in enhancing and expanding machine learning applications. Recently, these two technologies together gave birth to a new technology which is known as intelligent Cloud [15].

Cloud Computing (CC) [2] is one of the easy, flexible and quickly growing and most demanded technologies meant for delivering services requested by the users on demand over the Internet. The constraints such as cost, computational processing power, storage, analysis etc. involved in traditional approach have led to the raise of cloud computing [16]. It allows us to access various applications and data remotely without letting us install any software’s explicitly in personal laptops. The various services provided by CC “are generally categorized into Software as a Service (SaaS), Platform as a Service (PaaS) and Infrastructure as a Service (IaaS)” [2].

Cloud computing provides all the resources required to develop, run and deploy machine learning models on demand. Machine learning needs huge amount of data storage, computing power and many servers to concurrently work on models as presented in Fig. 1.

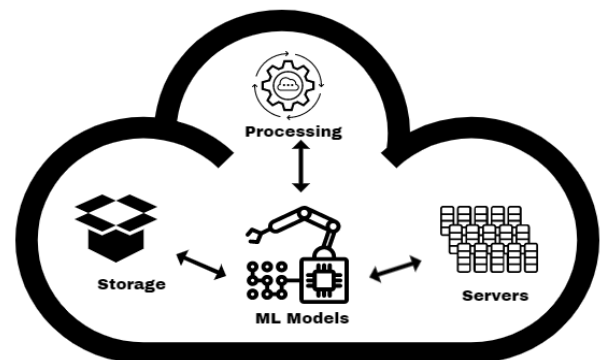


Fig. 1. Cloud Computing and Machine Learning.

Some of the key capabilities behind cloud with machine learning is that cloud’s pay per use service which is good for organizations who aspire to influence machine learning competences for their business without much spending. It offers the elasticity to work with machine learning features without having advanced data science skills. It helps us in ease of testing several machine learning skills and scales up as projects go into production and demand rises. Due to these capabilities, many cloud service providers today are offering lots of machine learning services for everyone without having background expertise of Artificial Intelligence (AI) and Machine Learning (ML).

Metropolitan cities today are loaded with huge population impacting solid waste growth such as food waste, plastics, bottles, sanitary waste, construction waste etc. impacting our surrounding environment. To minimize the effect of waste growth, it is necessary to understand and analyze the speed at which solid waste is being created. Existing waste management systems do not have automated techniques incorporated for exact prediction of solid waste growth [5]. Due to lack of data, incomplete data and other challenges such as poor strategies, they are not performing efficiently. To overcome this, machine learning approach can be used [1].

## II. EASE OF USE

Today's municipal waste management systems [8] are inefficient to perform waste analysis and take precautionary measures due to numerous loop holes such as lack of data, lack of technical expertise, lack of efficient strategies, lack of planning etc. Inaccurate prediction may be the reason for well-known shortfalls in waste administration arrangement such as unnecessary or inadequate disposal arrangement, waste collection, landfilling and recycling divisions. Accurate forecast is very much needed in case of metropolitan cities like Bengaluru, New Delhi etc. in India as they are highly populated impacting waste growth so that an appropriate action can be taken prior. These actions are not only to develop and improve existing systems but also help to alert the public so as to encourage decrease of waste and also recycle the solid waste produced. If the waste generated is not handled well, it may affect environment and living organisms' health. Due to noteworthy influence of waste growth on the environment, waste management systems [12] resulting minimal impacts on universe and zones required to be established. "Various methods of forecasting solid waste growth [6] can be generally categorized into five key clusters: descriptive statistical approach, regression approach, material flow approach, time series approach [13] and artificial intelligence approach" [5].

In [20], authors have used ANN for forecasting waste growth in Poland. Various explanatory variables were used to reveal the impact of socio-economic and demographic variables on the amount of waste generated. Performance of the models are measured using MSE (Mean Squared Error) and  $R^2$  metrics. The results proved that ANN is cost efficient approach in foreseeing the waste growth.

In [21], authors have developed hybrid Multilayer Perceptron (MLP) deep learning automated method to classify the waste dumped by community in the metropolitan area. Their experiments employed camera to capture images of waste and sensors to recognize the essential features. The experimental results proved that hybrid approach is capable to achieve more than 90% accuracy.

In [24], authors have carried out their research work to foresee waste growth for Mashhad city for different seasons using ANN on time series data. In [23], authors used weekly time series data to predict waste growth in Mashhad using SVM along with PCA (Principal Component Analysis). Kumar et al. used time series data which holds the yearly MSW (Municipal Solid Waste) [14] produced in New Delhi, India. Different models are used to predict waste growth and

the model's performance is assessed using RMSE and the IA values.

In [17], authors presented ARIMA model to foresee solid waste growth for Arusha city, Tanzania. Monthly generated waste data for the last few years 2008 to 2013 was used to carry out the research. The result proved that ARIMA (1, 1, 1) is well suited for forecasting "in terms of MAPE, MAD and RMSE measures".

In [18], authors presented "ARIMA model to forecast solid waste growth in the Kumasi Metropolitan Assembly (KMA)". The results showed that ARIMA (1, 1, 1) is well suited for predicting solid waste growth in the KMA.

In [19], authors presented "ARIMA model to forecast healthcare waste growth for the hospitals of Garhwal region of Uttarakhand, India". The performance of the model was analyzed using  $R^2$  value, MSE and MAE metrics and proved that ARIMA is best suited for forecast.

In [22], authors developed "ARIMA model to forecast the municipal solid waste growth of Abuja city, Nigeria. The results proved that an ARIMA (1, 1, 9) is the optimal model for forecast".

In [25], authors developed ARIMA model for forecasting amount of solid waste growth for Karur town, Tamil Nadu. Monthly based historical data was used for the year 2015 to 2017 and the results proved that ARIMA is best for prediction.

In [26], authors presented "ARIMA, Support Vector Regression (SVR) [4, 11], Grey model and Linear Regression (LR) model to forecast medical waste growth of Istanbul city, Turkey". Historical dataset used for forecast was from 1995 to 2017. Various performance metrics such as MAD, MAPE, RMSE and  $R^2$  were used to assess the models performance. Outcome showed that ARIMA (0, 1, 2) model is well suited for waste growth forecast.

## III. MACHINE LEARNING TIME SERIES FORECAST MODEL

Some of the key challenges in Statistics and Data Science today are time series and forecasting. A data is said to be time series data when it is bounded to time like days, months and years. When this data is used to predict future values, then it is called as Time series data.

ARIMA [3] is one of the popularly used machine learning algorithm for time series forecasting. It predicts future values using past data (autoregressive, moving average).

### ARIMA Model

It is a class of linear models that uses historical data to estimate forthcoming values. ARIMA [10] enclosing three components, Auto Regressive (AR), Integrated (I) and the Moving Average (MA) contribute to the ultimate forecast. Two of the Key concepts behind these models are Stationarity and Autocorrelation.

Stationarity tells that observations/data are time independent whereas autocorrelation relates the same set of observations but across diverse timing. The different components of ARIMA are explained as follows.

### Auto Regressive (AR)

This component uses autocorrelation concept, where the dependent features depend on the past values.

The general equation is:

$$X_t = \alpha_1 + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \dots + \phi_n X_{t-n} \quad (1)$$

As shown in (1), an observation  $X$  at time  $t$ ,  $X_t$ , depends on  $X_{t-1}, X_{t-2}, \dots, X_{t-n}$ ,  $\phi_1, \phi_2, \dots, \phi_n$  determines coefficient of lags that the model evaluates,  $\alpha_1$  is the intercept term, and where  $n$  is called the lag order which represents the number of previous lag samples or observations to be considered by the model.

### Integrated (I)

This component of ARIMA transforms non-stationary time-series data to a stationary by accomplishing prediction on the difference between any two pair of observations instead directly on the data itself.

$$\begin{aligned} A_t &= M_{t+1} - M_t \dots c = 1 \\ B_t &= A_{t+1} - A_t \dots c = 2 \end{aligned} \quad (2)$$

As shown in (2), Differencing tasks which can be achieved many times levels ( $M \rightarrow A$  and  $A \rightarrow B$ ), depends on the hyper parameter  $c$  that is set while training the ARIMA model.

### Moving Average(MA)

This component performs some kind of aggregation on the historical time series data in terms of residual error epsilon ( $\epsilon$ ) thus reducing noise in the data.

$$X_t = \alpha_2 + \omega_1 \epsilon_{t-1} + \omega_2 \epsilon_{t-2} + \dots + \omega_n \epsilon_{t-n} + \epsilon_t \quad (3)$$

The terms  $\epsilon$  indicate the residual errors from the aggregation operation as shown in (3) and  $n$  is another hyper parameter that specifies the time window for the moving average's residual error.  $X_t$  depends on the lagged forecast errors.

Following is the generic steps followed for ARIMA.

- Step 1: Visualization of Time Series Data
- Step 2: If data is non stationary, then convert it to stationary
- Step 3: Make the Correlation and AutoCorrelation graphs
- Step 4: Build the model using data
- Step 5: Make predictions using the model

## IV. RESULTS AND DISCUSSIONS

The ARIMA model used here for waste growth forecast is implemented in python. Jupyter notebook from Google Colab which is a free cloud service is used for the implementation.

Autocorrelation (AC) and Partial Autocorrelation (PAC) graphs shown in Fig. 2 (a) and (b) respectively are used to analyse and forecast future waste growth. They basically indicate how many days of previous data need to be considered to forecast future values which is known as lags. To calculate AR, three values,  $p$ ,  $d$ ,  $q$  need to be chosen,

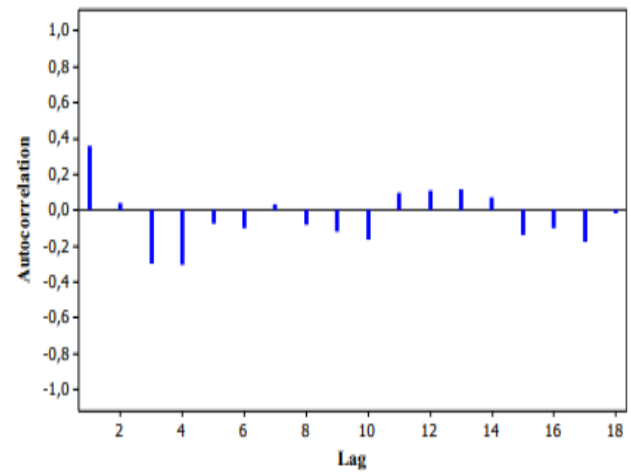
where,  $p$  represents AR model lags,  $d$  represents Differencing,  $q$  represents MA lags.

ARIMA model used here is to predict future trends of waste growth. The model needs stationary data to determine AR and MA components. Since the data used in this research work is non stationary,  $d=1$ , first order differencing was done.

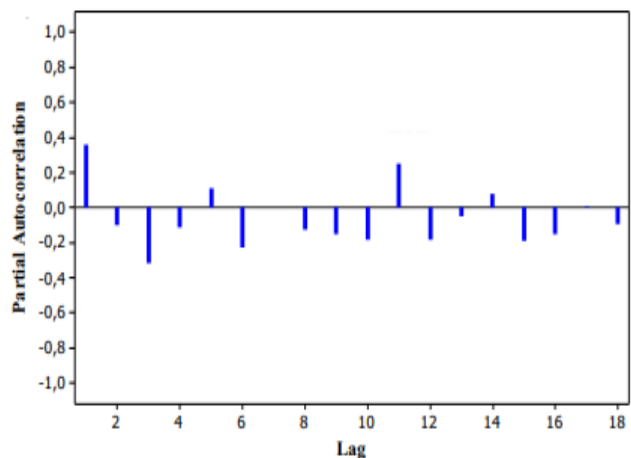
Autocorrelation graph in Fig. 2(a) showed that the series is stationary after first differencing. The arrangements of the AC and PAC graphs of the differenced series were examined for the initial computation of autoregressive ( $p$ ) and moving average orders ( $q$ ) in ARMA ( $p, q$ ) model.

For an AR model, the number of nonzero partial autocorrelations gives the most extreme lag of  $x$  that is used as a predictor.

Once, AC and PAC computation and analysis was done, ARIMA model was invoked on the dataset which gave the results shown in Table I and the forecast graph obtained is shown in Fig. 3.



(a)



(b)

Fig. 2. (a) Autocorrelation and (b) Partial Autocorrelation after First Differencing .



Equations used to measure the performance of the model are RMSE shown in (4), MAD shown in (5), MAPE% shown in (6), and R<sup>2</sup> shown in (7).

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (W_{real_i} - W_{expected_i})^2}{n}} \quad (4)$$

$$MAD = \frac{1}{n} \sum_{i=1}^n |W_{real_i} - W_{expected_i}| \quad (5)$$

$$MAPE = \frac{1}{n} \sum_{i=1}^n \frac{|W_{real_i} - W_{expected_i}|}{|W_{real_i}|} \times 100\% \quad (6)$$

$$R^2 = \left[ \frac{\sum_{i=1}^n (W_{real_i} - \overline{W_{real_i}})(W_{expected_i} - \overline{W_{expected_i}})}{\sqrt{\sum_{i=1}^n (W_{real_i} - \overline{W_{real_i}})^2 \times \sum_{i=1}^n (W_{expected_i} - \overline{W_{expected_i}})^2}} \right]^2 \quad (7)$$

Where,  $W_{real_i}$  and  $W_{expected_i}$  denote the real and expected value of  $i^{th}$  data point value, respectively.  $\overline{W_{real_i}}$  and  $\overline{W_{expected_i}}$  are the average of the real and expected value of  $i^{th}$  data point value. Also,  $n$  indicates the total number of data values. The performance of the model was measured by computing  $R^2$  value. It accepts values between 0 and 1, and values very close to 1 which indicates better fitting.

TABLE I. ARIMA MODEL PERFORMANCE

Model	RMSE	MAD	MAPE	R <sup>2</sup>
ARIMA (1,0,2)	854.2914	635.4722	11.5771	0.9767
ARIMA (0,1,0)	960.1165	713.0000	13.5741	0.9690
ARIMA (1,2,1)	1045.2257	777.5254	15.3163	0.9683
ARIMA (0,1,2)	753.5742	577.4601	11.6484	0.9788

Various ARIMA models are also made for selecting the model and their performance analysis is done using various metrics. As shown in Table I, the ARIMA (0, 1, 2) model has the highest R<sup>2</sup> (0.9788) and lowest RMSE (753.5742), MAD (577.4601), and MAPE (11.6484) and hence it is chosen as best model.

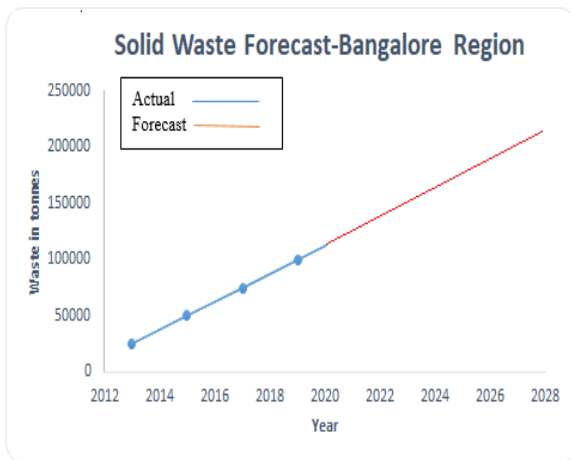


Fig. 3. Solid Waste Growth Forecast using ARIMA.

## V. CONCLUSION

It is very important to contribute and enhance the existing condition and scenarios of waste management in the crowded smart city, Bengaluru which can only be attained with precise waste assessment. Hence, the goal of this research work is to deliver an appropriate model to assess the quantity of waste produced. In this context, ARIMA (0, 1, 2) was chosen as the best model and used to forecast the waste growth of Bengaluru based on eight years of historical data. The outcomes of this research work can help waste management authorities to develop a reliable waste forecast model, which can be a significant foundation of information for Bengaluru. In addition, previous data about the volume of waste produced can be used for both the planning and design of future services.

## VI. LIMITATIONS AND FUTURE RESEARCH WORK

This research work targets to the development and improvement of waste management practices in smart cities through forecasting waste generation. It shows the development of a systematic process where time based factors affecting waste generation in smart cities have been determined to study and forecast waste growth. Unavailability of the continuous waste data and also socio-economic and demographic variables affecting solid waste generation makes it difficult to foresee solid waste growth for the developing countries like India.

The research work can be extended in the future by incorporating more input features, more socio-economic parameters. Other ML, AI or deep learning techniques can be used in future to handle complex scenarios and to achieve better accuracy.

## REFERENCES

- [1] X. Cuong Nguyen, T. Thanh Huyen Nguyen, D. Duong La, Gopalakrishnan Kumar, Eldon R. Rene, D. Duc Nguyen, S. Woong Chang, W. Jin Chung, X. Hoan Nguyen, V. Khanh Nguyen, Development of machine learning - based models to forecast solid waste generation in residential areas: A case study from Vietnam, Resources, Conservation and Recycling, Volume 167, 2021, 105381, ISSN 0921-3449, <https://doi.org/10.1016/j.resconrec.2020.105381>.
- [2] Butt, Umer Ahmed, Muhammad Mehmood, Syed Bilal Hussain Shah, Rashid Amin, M. Waqas Shaikat, Syed Mohsan Raza, Doug Young Suh, and Md. Jalil Piran. 2020. "A Review of Machine Learning Algorithms for Cloud Computing Security" *Electronics* 9, no. 9: 1379. <https://doi.org/10.3390/electronics9091379>Ceylan Z, Bulkan S, Elevli S. Prediction of medical waste generation using SVR, GM (1,1) and ARIMA models: a case study for megacity Istanbul. *J Environ Health Sci Eng.* 2020 Jun 19;18(2):687-697.
- [3] Ş. T. Özçelik and F. Boray Tek, "Forecasting and Analysis of Domestic Solid Waste Generation in Districts of Istanbul with Support Vector Regression," *2020 5th International Conference on Computer Science and Engineering (UBMK)*, Diyarbakir, Turkey, 2020, pp. 366-371. doi: 10.1109/UBMK50275.2020.9219368.
- [4] Soni, U., Roy, A., Verma, A. *et al.* Forecasting municipal solid waste generation using artificial intelligence models—a case study in India. *SN Appl. Sci.* 1, 162 (2019). <https://doi.org/10.1007/s42452-018-0157-x>
- [5] D. M. S. H. Dissanayaka and S. Vasanthapriyan, "Forecast Municipal Solid Waste Generation in Sri Lanka," *2019 International Conference on Advancements in Computing (ICAC)*, Malabe, Sri Lanka, 2019, pp. 210-215, doi: 10.1109/ICAC49085.2019.9103421.
- [6] Miyuru Kannagara, Rahul Dua, Leila Ahmadi, Farid Bensebaa, Modeling and prediction of regional municipal solid waste generation

- and diversion in Canada using machine learning approaches, *Waste Management*, Volume 74, 2018, pp. 3-15, ISSN 0956-053X.
- [7] Maya Chavan, T. R. Pattanshetti, "Survey on Municipal Waste Collection Management in Smart City", *International Research Journal of Engineering and Technology (IRJET)*, Volume: 05 Issue: 01, Jan-2018, p-ISSN: 2395-0072 e-ISSN: 2395-0056.
- [8] M. Talha, A. Upadhyay, R. Shamim and M. S. Beg, "A cloud integrated wireless garbage management system for smart cities," *2017 International Conference on Multimedia, Signal Processing and Communication Technologies (IMPACT)*, Aligarh, 2017, pp. 175-179, doi: 10.1109/MSPCT.2017.8363999.
- [9] Chauhan, Ankur & Singh, Amol. (2017). An ARIMA model for the forecasting of healthcare waste generation in the Garhwal region of Uttarakhand, India. *International Journal of Services Operations and Informatics*. 8. 10.1504/IJSOI.2017.086587.
- [10] C. Dai, Y.P. Li, G.H. Huang, A two-stage support-vector-regression optimization model for municipal solid waste management – A case study of Beijing, China. *Journal of Environmental Management*, Volume 92, Issue 12, 2011, pp. 3023-3037.
- [11] S. Lebersorger, P. Beigl, Municipal solid waste generation in municipalities: Quantifying impacts of household structure, commercial waste and domestic fuel, *Waste Management*, Volume 31, Issues 9–10, 2011, pp. 1907-1915, ISSN 0956-053X, <https://doi.org/10.1016/j.wasman.2011.05.016>.
- [12] Mwenda A, Kuznetsov D, Mirau S, Time series forecasting of solid waste generation in Arusha city- Tanzania. *Mathematical Theory and Modeling*, Vol.4, No.8, 2014, pp. 29-39, ISSN 2224-5804 (Paper) ISSN 2225-0522 (Online).
- [13] Hanandeh, Ali & Abbasi, Mariam. (2016). Forecasting municipal solid waste generation using artificial intelligence modelling approaches. *Waste Management*. 56. 10.1016/j.wasman.2016.05.018.
- [14] Aazam, Mohammad et al. "Cloud-based smart waste management for smart cities." *2016 IEEE 21st International Workshop on Computer Aided Modelling and Design of Communication Links and Networks (CAMAD)* (2016): pp. 188-193.
- [15] C. Ji, Y. Li, W. Qiu, U. Awada and K. Li, "Big Data Processing in Cloud Computing Environments," *2012 12th International Symposium on Pervasive Systems, Algorithms and Networks*, San Marcos, TX, 2012, pp. 17-23.
- [16] Amon Mwenda, Dmitry Kuznetsov, Silas Mirau, "Time Series Forecasting of Solid Waste Generation in Arusha City – Tanzania", *Mathematical Theory and Modeling* www.iiste.org ISSN 2224-5804 (Paper) ISSN 2225-0522 (Online) Vol.4, No.8, 2014.
- [17] Owusu-Sekyere, Ebenezer. (2013). Forecasting and planning for solid waste generation in the Kumasi metropolitan area of Ghana: An ARIMA time series approach. *International Journal of Sciences*. 2. 69-83.
- [18] Chauhan, Ankur & Singh, Amol. (2017). An ARIMA model for the forecasting of healthcare waste generation in the Garhwal region of Uttarakhand, India. *International Journal of Services Operations and Informatics*. September 13, 2017pp 352-366 8. 10.1504/IJSOI.2017.086587.
- [19] Kulisz, M.; Kujawska, J. Prediction of municipal waste generation in Poland using neural network modeling. *Sustainability* 2020, 12, 10088.
- [20] Chu, Yinghao & Huang, Chen & Xie, Xiaodan & Tan, Bohai & Kamal, Shyam & Xiong, Xiaogang. (2018). Multilayer Hybrid Deep-Learning Method for Waste Classification and Recycling. *Computational Intelligence and Neuroscience*. 2018. 1-9. 10.1155/2018/5060857.
- [21] U. A. Dodo, E. C. Ashigwuike and J. N. Emechebe, "Municipal Solid Waste Generation Forecast using an ARIMA Model: A Focus on Abuja City, Nigeria," *2022 IEEE Nigeria 4th International Conference on Disruptive Technologies for Sustainable Development (NIGERCON)*, 2022, pp. 1-5, doi: 10.1109/NIGERCON54645.2022.9803108.
- [22] Noori, R, Khakpour, A, Omidvar, B, et al. (2010b) Comparison of ANN and principal component analysis-multivariate linear regression models for predicting the river flow based on developed discrepancy ratio statistic. *Expert Systems with Applications* 37: 5856–5862.
- [23] Jalili, Mojtaba & Noori, Rahimullah. (2007). Prediction of Municipal Solid Waste Generation by Use of Artificial Neural Network: A Case Study of Mashhad. *International Journal of Environmental Research* (ISSN: 1735-6865) Vol 2 Num 1. 2.
- [24] Vignesh.C, Rameshkumar.M, Dr.S.Anand Kumar Varma, Time Series Forecasting of Solid Waste Generation in Karur City -Tamil Nadu, *SSRG International Journal of Civil Engineering (SSRG - IJCE)* – Volume 5 Issue 2 - February 2018 ISSN: 2348 – 8352 .
- [25] Ceylan, Z., Bulkan, S. & Elevli, S. Prediction of medical waste generation using SVR, GM (1,1) and ARIMA models: a case study for megacity Istanbul. *J Environ Health Sci Engineer* 18, 687–697 (2020). <https://doi.org/10.1007/s40201-020-00495-8>.

# Building an Intelligent Tutoring System for Learning Polysemous Words in *Mooré*

Pengwendé ZONGO

Laboratoire Mathématiques, Informatique et Applications  
Université Norbert ZONGO  
Koudougou, Burkina Faso

Tounwendyam Frédéric OUEDRAOGO

Laboratoire Mathématiques, Informatique et Applications  
Université Norbert ZONGO  
Koudougou, Burkina Faso

**Abstract**—This paper presents the results of our research carried out as part of the building of an Intelligent Tutoring System (ITS) to learn *Mooré*, a tone language. A word in tone language may have many meanings according to the pitch. The system has an intelligent tutor to personalize and guide the learning of the transcription of polysemous words in *Mooré*. This learning activity aims both to master the transcription and also to distinguish the lexical meaning of words according to the pitch used. A first step of this research has been the specification of the processes, inference and knowledge of the system. In this work we present the implementation and pedagogical assessment of the system. We designed the architecture of the ITS, the diagnosis of transcription errors and remediation approach. Then, we used the Petri net formalism to model the system dynamic in order to analyze its states and fix deadlocks. We developed the system in java and we evaluated its educational value by an experimentation with learners. This shows that the learning objectives can be achieved with this system.

**Keywords**—Intelligent tutoring system; petri network; evaluation; *Mooré* language

## I. INTRODUCTION

An Intelligent Tutorial System (ITS) is a computing environment for human learning that integrates artificial intelligence techniques and cognitive theories in order to provide guided and personalized learning to learners [1], [2]. It consists of four main modules: domain, student, tutoring and communication [26], [27]. In the literature, ITS for learning language research are mostly on European and Asian languages such as English, Japanese, Chinese [5], [10], [11], [12]. Our research contributes to the development of ITS for language learning. We aim to build an ITS to learn polysemous words in *Mooré* through transcription activities. *Mooré* is a tone language, the most spoken in Burkina Faso. This language is also spoken in some neighboring countries such as Côte d'Ivoire, Ghana, Mali and Togo.

The contribution presented in this paper follows a previous one where we presented a specification of an intelligent tutoring system to learn tone language [13]. We used CommonKADS a knowledge engineering method to specify the knowledge and processes of the system. The specification provides a common framework for the development of transcription-based ITS for any tone language.

In Burkina Faso, the learning of local languages is part of the non-formal education program of the government. So, only with the training centers, often of short-term projects, provide learning programs for local languages. Therefore, building

an ITS for *Mooré* learning is a significant contribution in the field of local languages ilearning in Burkina Faso. The building of such IT tools could not only help meet the needs to learn local languages in sub-Saharan Africa but also to ensure the continuity of local language learning during periods of pandemic such as COVID-19.

The rest of the paper is organized as follows. Section II presents the background. It contains important concepts used in this article. In Section III we present the architecture of the system and describe the approach to diagnostic transcription errors and the remediation to provide. Section IV shows the development framework and an overview of the system. Section V presents the evaluation results of the system experimentation. In Section VI we summarize the work done and gives some perspectives.

## II. BACKGROUND

In this section, we present important concepts that we used and related work on tone language. These concepts are the Petri net, Bayesian network widely used in the field of ITS research.

### A. Tone Languages

A Tone language includes pitch phonemes in addition to consonants and vowels, and pitch differences are used to distinguish one lexical item from another [15]. Tone languages are characterized by two types of tones: punctual tone and melodic tone. In punctual tones, only one aspect of the melodic curve is considered (highest or lowest) whereas in modulated tones, they are distinguished by successive directions of the melodic curve [16]. *Mooré* and most of the languages spoken by sub-Saharan African are languages with punctual tones [4]. The *Mooré* language has three tonal patterns:

- the high tone, represented by the acute accent ( )
- the medium tone, represented by the dash sign ( )
- the low tone, represented by the grave accent ( )

The Non-respect of tones leads to confusion, misinterpretation or nonsense.

The learning activities of the ITS are based on transcription tasks of polysemous words in *Mooré* language at the example of Fig. 1. To do this, taking account the tone in the transcription is very important because it allows us to distinguish the lexical meaning of words according to the pitch.

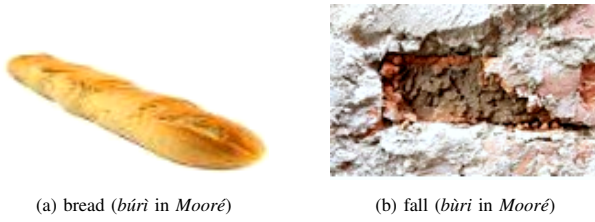


Fig. 1. Example of Two Words(Bread and Fall) Represented in Images. The First Image is Transcribed by the Word *Bûri* and the Second Image by the Word *Bûri*. The Word *Bûri* in *Mooré* without the Pitch, it' is Unclear whether that Word Alludes to the First Image or to the Second Image.

### B. Petri Networks

The Petri networks in short Petri nets is a mathematical modeling formalism introduced by Dr. Carl Adam Petri in 1962. This modeling tool is used to represent the dynamics of discrete distributed systems in computer science, engineering and so forth [8]. The petri nets formalism is borrowed from graph theory. Therefore, a Petri nets is a directed bipartite graph that has two types of node: places and transitions. Place is represented by circle and transition by bar or box. Places and transitions are connected by directed edges. Place represents system states, condition or resources that must be met before an action can be performed. Transition represents actions.

A mathematical definition of Petri net is a tuple  $PN = (P, T, Pre, Post)$  where:

- $P = \{p_1, p_2, \dots, p_n\}, n > 0$  a finite set of places;
- $T = \{t_1, t_2, \dots, t_m\}, m > 0$  a finite set of transitions;
- The places  $P$  and transitions  $T$  are disjoint ( $P \cap T = \emptyset$ );
- $Pre : (PxT) \rightarrow \mathbb{N}$  is an input function that defines directed edges from places to transitions;
- $Post : (TxP) \rightarrow \mathbb{N}$  is an output function that defines directed edges from transitions to places.

A marked Petri net is a five tuple  $G = (P, T, Pre, Post, M)$  where  $M$  can be viewed as a function, which assigns a natural number with each place, i.e.  $M : P \rightarrow \mathbb{N}$ .  $M$  can also be viewed as a vector given by  $M_k = \{M_1, M_2, \dots, M_i, \dots, M_n\}$  where the  $i^{th}$  entry of  $M$  is  $M_i$ , which is the marking of the place  $p_i$ . The execution of a Petri net causes its marking to change by removing tokens from its input places and depositing into each of its output places.

A transition is said to be enabled when each one of its input places is marked with at least one token. In mathematical terms, a transition,  $t \in T$ , is enabled if  $M(p) \geq Pre(p, t); \forall p \in P$ . If an enabled transition  $t$  fires then it causes a change in marking from  $M(p)$  to  $M'(p)$  given by the equation:

$$M'(p) = M(p) - Pre(p, t) + Post(p, t); \forall p \in P.$$

As for the usefulness of the Petri net, this approach is used to diagnose modeling errors of an application [9], [6]. In the field of ITS, Petri nets have been used to model systems

and to verify their consistency [7]. Thus, for our study, we used graphical representation of Petri net to model the operations(actions) and states of our system. The reachability graph will allow to represent the different firings of the marked Petri net. The purpose of the reachability graph is to remove possible deadlock states of the system.

### C. Bayesian Network

The Bayesian network is a graphical and probabilistic model for representing uncertain or incomplete knowledge of the learner in the field of learning [3]. It is a technique of artificial intelligence initiated by Corbett and Anderson in 1994. The graphical model is represented by two parameters  $\{N, A\}$ .  $N$  represents the set of nodes or vertices and  $A$  the set of arcs. The probabilistic representation of the Bayesian network is based on Bayes' theorem [28]. Bayes' theorem is translated by the following mathematical formula:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

In other words, this equation means: for two events  $A$  and  $B$ , what is the probability that  $A$  will occur given  $B$ .

In the field of ITS, the Bayesian network approach can be used to [17]: update the learner model [18]; diagnose the causes of learner errors [19]; or predict the actions of the learner in a problem-solving process [20]. Diagnosing misconceptions requires collecting and checking for buggy rules, which sometimes leads to overwhelming and impractical numbers of buggy rules, even for simple domains such as fractions [21]. Modern approaches, such as algorithmic debugging [22], automatically distinguish buggy rules. With reference to the study on the specification of knowledge and processes, the authors have determined two inference structures in the context of the design of ITS for tone languages learning [13]. These two inference structures are: the inference structure for Assessment and the inference structure for Diagnosis. The following Task-decomposition diagram (see Fig. 2) presents the different inferences used by the inference structure for Assessment.

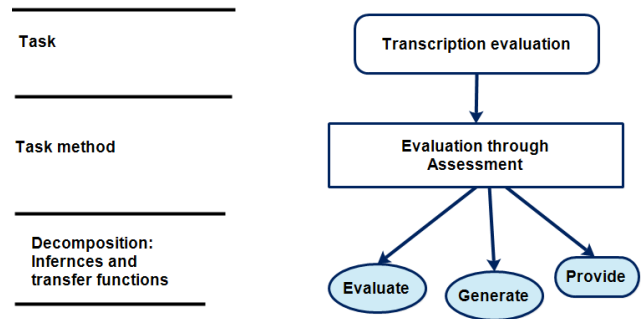


Fig. 2. Task-Decomposition Diagram.

For this present study, we use the Bayesian network approach to trace transcription errors. This is a very suitable approach for the diagnosis of cognitive knowledge.

### III. SYSTEM DESIGN

In this section of our study, we first describe the different components of the intelligent tutoring system to learn *Mooré* polysemous word, then we show the approach used for the knowledge diagnosis and finally we present the approach used to design the operation of the system.

#### A. Architecture of the ITS

As most intelligent tutors, the intelligent tutoring system for learning transcription of polysemous words in *Mooré* language consists of four components namely the domain module, the student module, the tutoring module and the communication module. However, for the modeling of the components of our system, it was made taking into account the specificity of the domain of learning. Fig. 3 presents this architecture.

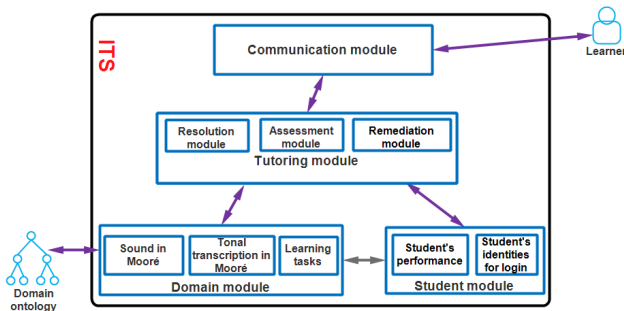


Fig. 3. Architecture of the System.

The different modules in Fig. 3 correspond to the following description:

- Domain module

The role of the domain module is to provide the system with all the information related to the knowledges of the learning area. The domain module consists of the learning tasks, transcriptions corresponding of learning tasks in *Mooré* and sounds in *Mooré* corresponding to the learning tasks. The Learning Tasks component of the domain module consists of all the learning tasks in the system. These tasks are presented in the form of images. The Tonal transcription in *Mooré* component contains the ideal transcriptions of the learning tasks. As for the Sound in *Mooré* component, it consists of sounds corresponding to ideal transcriptions in *Mooré*. The tutoring module uses this component of the domain module to provide didactic aid to learners. The didactic aid of our system allows learners to listen to the sound corresponding to the task selected in order to transcribe correctly.

- Tutoring module

The pedagogical strategy of the system is represented in the tutoring module. This module consists of the resolution, evaluation and remediation modules. With reference to the Petri network of our system represented by Fig. 6 in subsection III-C, the Resolution module is responsible for executing actions  $t_3$ ,  $t_4$  and  $t_5$  of the Petri network. As for the Assessment module, it is responsible for executing actions  $t_6$ ,  $t_7$ ,  $t_8$ ,  $t_9$  and

$t_{10}$ . For the Remediation module, it is responsible for executing actions  $t_{11}$ ,  $t_{12}$ ,  $t_{13}$ ,  $t_{14}$ ,  $t_{15}$ ,  $t_{16}$ ,  $t_{17}$ ,  $t_{18}$ ,  $t_{19}$  and  $t_{20}$  of the Petri network.

- Student module

It is composed of the learner's performance states in relation to his tasks solved and the data for the system login. The student module is responsible for managing all the information of learner's profile. It updates the profile of the latter, in particular the Student's performance component, after each resolution of a task.

- Communication module

The communication module consists of the different interaction windows between the system and the user. It allows the system to interact with users and vice versa.

We consider that the domain module and the tutoring module represent the most important modules of our ITS because domain module contains the knowledge that the system should taught and tutoring module the strategies that the system should use to evaluate learning and provide assistance.

#### B. Diagnostic

In the domain of ITS, some authors limit remediation to feedback [23], [24] and others perceive it as a process consisting of cognitive diagnosis and feedback [25]. We consider remediation as a process that consists of first detecting the sources of errors and then generating the appropriate feedback in relation to the error committed. So, for the diagnosis of knowledge, we have, using the graphic model of Bayesian network method, first proceeded to the representation of the uncertain and incomplete knowledge of the learners. This representation allowed us to make the causal links between this knowledge. We then developed the different algorithms for tracing transcription errors based on the Bayesian network representation. Fig. 4 below represents the Bayesian network model that we designed.

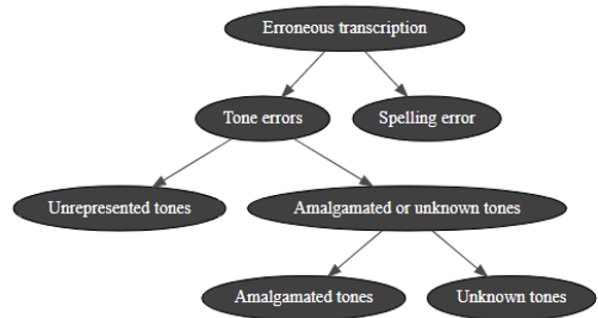


Fig. 4. Bayesian Network for Knowledge Diagnosis.

Fig. 4 presents the diagnosis of transcription errors. When the transcription evaluation made by the learner returns an erroneous transcription, the system performs tracing based on the above Bayesian network in order to detect the transcription error and generate the suitable feedback. The probable transcription errors listed in the graphical model of the Bayesian network above have been identified in collaboration with



the *Mooré* language trainers The flexibility of the Bayesian network approach allowed our collaboration non-computer scientists to easily understand the graphical model shown in Fig. 4. Also, this representation allowed the trainers of the *Mooré* language to participate in the validation of the study. Fig. 5 presents the flowchart model of our Bayesian network.

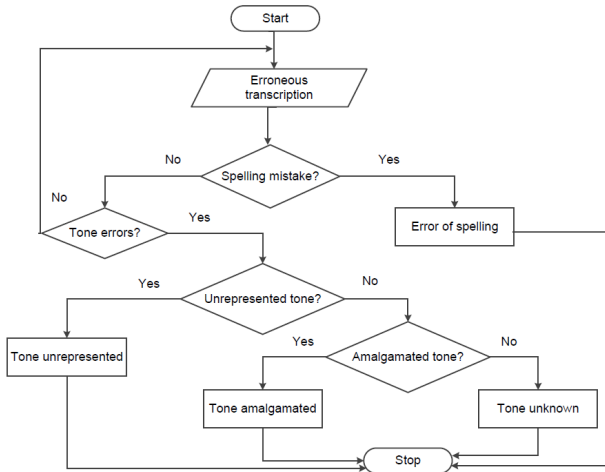


Fig. 5. Flowchart of Knowledge Diagnosis. This Represents the Bayesian Network Algorithm.

Fig. 5 shows a flowchart of the remediation algorithm that generates adaptive feedback according to the learner.

### C. Verification of the System Consistency

The Petri Network is an efficient tool for the verification of discrete event systems [14]. Since an ITS is a discrete event system, we use the Petri net formalism to model our system. This model is important to ensure that the system does not have any action or operation that would put it in a deadlock situation. Therefore, we can simulate the operation of the system and resolve possible blockage situations before implementation step.

Fig. 6 presents the Petri net of the system which models its different states.

In Fig. 6, the transitions ( $t_i$ ) represent the different actions performed by the system and the places ( $p_i$ ) represent the input or output data of the actions of system. Table I describes the different actions and states of the system.

Based on the description of places and transitions in Table I, the task solving, for example, is described as follows. The system first executes the action ( $t_3$ ). The learner selects one of the tasks presented ( $p_4$ ) and the system then executes the action ( $t_4$ ). From the data ( $p_5$ ), the system finally executes the action ( $t_5$ ).

To correct the possible blockages of the system, we made the reachability graph in order to analyze the different firings of the transitions. Fig. 7 presents the reachability graph of the Petri net.

The analysis of Fig. 7 shows that for each firing of  $t_i$ , the input place  $p_i$  goes from 1 to 0 and the output place  $p_{i+1}$  goes

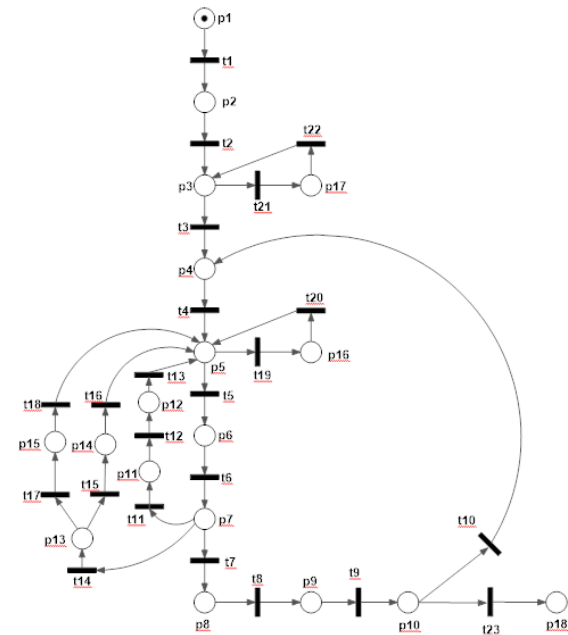


Fig. 6. Petri Net Representation of the System States (Places) and Actions (Transitions).

TABLE I. DESCRIPTION OF THE PLACES (STATES) AND TRANSITIONS (ACTIONS) OF THE PETRI NET.

Places ( $p_i$ )	Transitions ( $t_i$ )
$p_1$ : app icon	$t_1$ : display login screen
$p_2$ : login screen	$t_2$ : check login and password
$p_3$ : home screen(Main)	$t_3$ : load tasks
$p_4$ : tasks presented	$t_4$ : load image and audio
$p_5$ : task selected, image and audio loaded	$t_5$ : read transcription
$p_6$ : transcribed word read	$t_7$ : produce success feedback
$p_7$ : transcribed word assessed	$t_8$ : mark task solved
$p_8$ : success feedback generated	$t_9$ : update learner's profile
$p_9$ : task solved marked	$t_{10}$ : return to tasks presented
$p_{10}$ : learner's profile updated	$t_{11}$ : detect spelling mistake
$p_{11}$ : spelling mistake detected	$t_{12}$ : display spelling mistake
$p_{12}$ : feedback spelling mistake displayed	$t_{13}$ : transcribe again
$p_{13}$ : tones error detected	$t_{14}$ : detect tones error
$p_{14}$ : feedback amalgamated tones displayed	$t_{15}$ : detect amalgamated tones and produce feedback
$p_{15}$ : feedback tones error displayed	$t_{16}$ : transcribe again
$p_{16}$ : sound emitted	$t_{17}$ : produce tones error feedback
$p_{17}$ : learner score displayed	$t_{18}$ : transcribe again
$p_{18}$ : system ended	$t_{19}$ : emit sound
	$t_{20}$ : end sound emitted
	$t_{21}$ : display score
	$t_{22}$ : return to the main menu
	$t_{23}$ : stop the system

from 0 to 1. From these results, we can say that the Petri net is 1-safe (or binary) which means that the system is deadlock-free. We can conclude that the designed system is coherent.

## IV. SYSTEM OVERVIEW

We implemented the system on the Java environment. We present in this section an overview of the system functionalities. An user can use this application to learn *Mooré* language by transcription tasks. Fig. 8 presents the system dashboard view.

Fig. 8 presents the system dashboard which provides main menu. A click on the "TACHES" button leads to the resolution



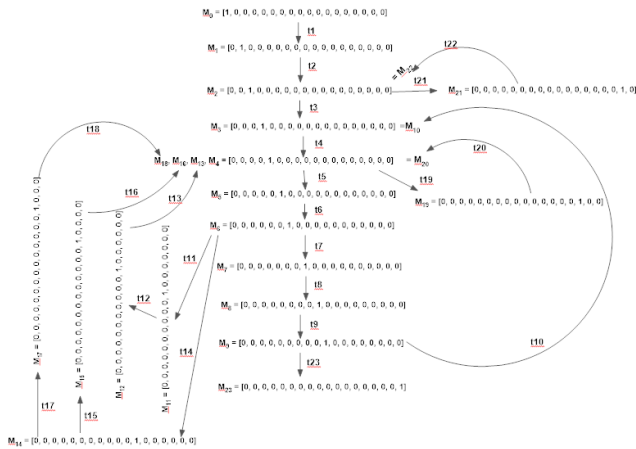


Fig. 7. Reachability Graph Corresponding to the Petri Net of Figure 6.



Fig. 10. An Example of Transcription. The Task Number 2 of Broom Image is Transcribed in *Mooré* by *saaga*.



Fig. 8. System Dashboard.

interface and user can select a task to transcribe, see Fig. 9.

The resolution interface displays a list of tasks to be solved by the learner. In Fig. 9, the selected task (number 2) is displayed as an image to be transcribed in *mooré*.

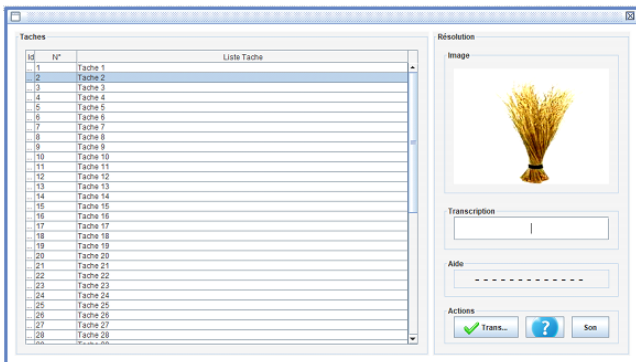
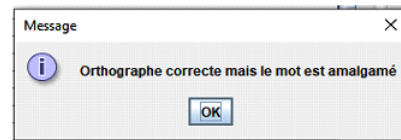


Fig. 9. Resolution Interface Presenting the Tasks. Each Task Corresponds to an Image to be Transcribed.

Fig. 10 gives an illustration of a resolution of task. The selected task shown in Fig. 9 consist of an image of broom. We suppose the user entered *saaga* in *Mooré* as the answer. We use this wrong answer that will allows to show some feedback generated by the system.



(a) Feedback of the Amalgamate

Fig. 11. Feedback Generated by the System when the User Enters the Word *saaga* as Answer. In this Case the Answer is Wrong Even if the Spelling is Correct.

The correct answer should be *saagà* with low tone on the last vowel.

## V. PEDAGOGICAL ASSESSMENT

To do the pedagogical assessment of our system, we proceeded with the experimentation of the application. A total of four *Mooré* trainers and seventeen learners were able to do the experimentation. For the questionnaires, we developed them via Google Forms.

The formulation of the questions addressed to the learners and the trainers aimed to verify the following aspects:

- The conformity between the content of the Knowledge Base (KB) of the system developed and the content of

the corpus of *Mooré* language, namely, the transcriptions and the sounds:

- The relevance of the learning tasks.
- The clarity of the resolution steps.
- The relevance of the feedback generated.
- The relevance of the sounds loaded.
- The ease of use of the system.
- And the contribution of the system in the field of the *Mooré* language learning.

Fig. 12 and Fig. 13 give an overview of some results of the users' feelings after the experimentation of the application.

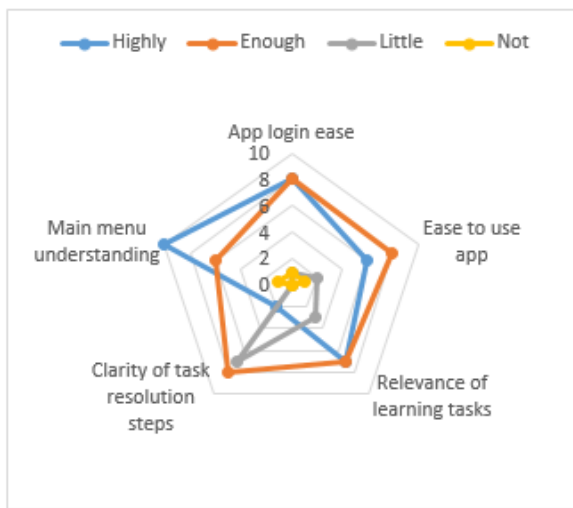


Fig. 12. Learners' Opinions on the System

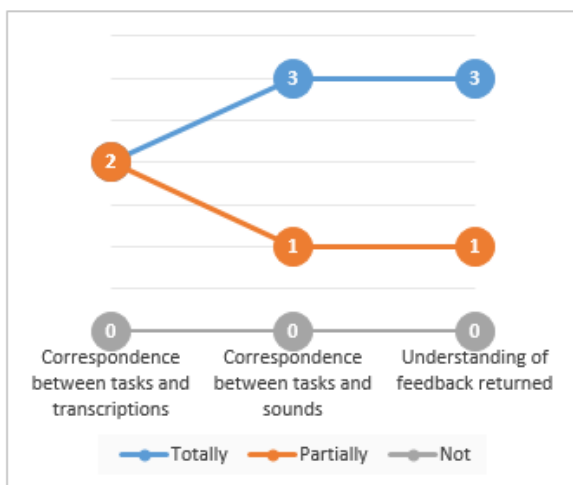


Fig. 13. Trainers' Opinions on the System.

We analysis the collection of the users' opinions and we present here some results in Fig. 12 and Fig. 13 for both learners and trainers.

For the verification of the conformity between the content of the Knowledge Base of the developed system and the content of the corpus of the *Mooré* language, the four trainers found an almost total conformity between the two contents. For the relevance of the learning tasks, fourteen out of seventeen learners found the different tasks at least enough relevant. For the clarity of the resolution steps, ten out of seventeen learners found the steps of resolution at least enough clear. As for the relevance of the feedback generated and the sounds loaded, fourteen learners affirmed that the feedback generated help in the correction of transcription errors and sixteen affirmed that the sounds emitted really help in transcription. And for the contribution of the system in the field of the *Mooré* language learning, all trainers answered in the affirmative that the system allows learning without human assistance.

Based on these results, we can say that the pedagogical assessment of the system allowed to show that the use of it could made it possible the learning of the transcription in *Mooré* without human assistance. Similarly, this system would be a great contribution in the field of education for local languages learning in Burkina Faso.

## VI. CONCLUSION AND PERSPECTIVES

In this paper, we presented the work on system design, the implementation of the system and the pedagogical assessment of the system. In the system design, we showed the architecture of the ITS, we described the Bayesian network approach that we used to trace the errors of transcription and we presented the Petri net approach that we used to design the operation of the system. The architecture proposed allowed to define the different modules of the system. The Bayesian network model made it possible to represent the uncertain knowledge and to develop the algorithms for tracing errors. As for the Petri net approach, it allowed to simulate the operation of the system and to correct the possible blockages. In relation to the implementation of the system, we presented some views of the system implemented. And as for the pedagogical assessment of the system, the experimentation of the application made it possible to collect the users' feelings. The analysis of these feelings shows, among other things, that our system is a great contribution for *Mooré* learning. For the future, we expect to develop a speech recognition activity to integrate into our system. This speech recognition activity will learn phonetics in *Mooré*. We also expect to develop a WordNet ontology for the *Mooré* language. The WordNet ontology for the *Mooré* language is a particularly promising avenue. It will constitute an online knowledge base and will be interoperable with our system and with other applications.

## REFERENCES

- [1] Arthur C Graesser, Mark W Conley, and Andrew Olney. Intelligent tutoring systems. APA educational psychology handbook, Vol 3: Application to learning and teaching., pages 451–473, 2012.
- [2] Indira Padayachee. Intelligent tutoring systems: Architecture and characteristics. In Proceedings of the 32nd Annual SACLA Conference, pages 1–8. Citeseer, 2002.
- [3] Catherine L. Caldwell-Harris, Alia Lancaster, D. Robert Ladd, Dan Dediu, Morten H. Christiansen, Factors Influencing Sensitivity To Lexical Tone In An Artificial Language. Implications for Second Language Learning, pp. 335 – 357, 2015.
- [4] Laetitia Compaoré, *Mooré* prosody analysis essay: tone and intonation. Linguistics. Doctoral thesis, Université de Sorbonne Paris, p. 19, 2017

- [5] J. Paladines and J. Ramirez, "A Systematic Literature Review of Intelligent Tutoring Systems With Dialogue in Natural Language," in *IEEE Access*, vol. 8, pp. 164246-164267, 2020, doi: 10.1109/ACCESS.2020.3021383.
- [6] Jiliang Luo, Qi Zhang, Xuekun Chen, and MengChu Zhou. Modeling and race detection of ladder diagrams via ordinary petri nets. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 48(7):1166–1176, 2017.
- [7] Yu-Ying Wang, Ah-Fur Lai, Rong-Kuan Shen, Cheng-Ying Yang, Victor RL Shen, and Ya-Hsuan Chu. Modeling and verification of an intelligent tutoring system based on petri net theory. *Mathematical Biosciences and Engineering*, 16(5):4947–4975, 2019.
- [8] Carl Adam Petri. *Kommunikation mit automaten*. 1962.
- [9] Feng Chu. Conception des systèmes de production à l'aide des réseaux de Petri: vérification incrémentale des propriétés qualitatives. PhD thesis, Université Paul Verlaine-Metz, 2015.
- [10] Abdelbaset Almasri, Adel Ahmed, Naser Al-Masri, Yousef Abu Sultan, Ahmed Y. Mahmoud, Ihab Zaqout, Alaa N. Akkila, Samy S. Abu-Naser, *Intelligent Tutoring Systems Survey for the Period 2000- 2018*, International Journal of Academic Engineering Research (IJAER) ISSN: 2000-003X, Vol. 3 Issue 5, p. 21-37, May – 2019
- [11] Ahuja, Neelu Jyothi and Sille, Roohi, A critical review of development of intelligent tutoring systems: Retrospect, present and prospect, *International Journal of Computer Science Issues (IJCSI)*, Vol. 10, Issue 4, No 2, pp. 39-48, July 2013
- [12] Slavuj, V., Kovačić, B., & Jugo, I. (2015, May). Intelligent tutoring systems for language learning. In 2015 38th International Convention on Information and Communication Technology, Electronics and Micro-electronics (MIPRO) (pp. 814-819). IEEE.
- [13] Pengwendé Zongo and T. Frédéric Ouedraogo. Toward An Intelligent Tutoring System for Tone Languages: learning of tone levels in Mooré. In 22nd IEEE International Conference on Advanced Learning Technologies (ICALT 2022).
- [14] Uzam, M. U. R. A. T., & Jones, A. H. (1998). Discrete event control system design using automation Petri nets and their ladder diagram implementation. *The International Journal of Advanced Manufacturing Technology*, 14(10), 716-728.
- [15] Caldwell-Harris, C. L., Lancaster, A., Ladd, D. R., Dediu, D., & Christiansen, M. H. (2015). Factors influencing sensitivity to lexical tone in an artificial language: Implications for second language learning. *Studies in Second Language Acquisition*, 37(2), 335-357.
- [16] Otsaga, T. A. (2002). Les tons dans les dictionnaires de langues gabonaises: situation et perspectives. *Lexikos*, 12.
- [17] Pelleu-Tchétagani, J. M. (2005). Une approche propéagogique du diagnostic cognitif dans les STI: conception, formalisation et implémentation.
- [18] Tchétagani, J. M., & Nkambou, R. (2002, June). Hierarchical representation and evaluation of the student in an intelligent tutoring system. In *International conference on intelligent tutoring systems* (pp. 708-717). Springer, Berlin, Heidelberg.
- [19] Tchétagani, J., Nkambou, R., & Kabanza, F. (2004, May). Epistemological remediation in intelligent tutoring systems. In *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems* (pp. 955-966). Springer, Berlin, Heidelberg.
- [20] Conati, C., Gertner, A., & Vanlehn, K. (2002). Using Bayesian networks to manage uncertainty in student modeling. *User modeling and user-adapted interaction*, 12(4), 371-417.
- [21] Jacqueline Bourdeau, Monique Grandbastien. La modélisation du tutorat dans les systèmes tutoriels intelligents. *STICEF (Sciences et Technologies de l'Information et de la Communication pour l'Éducation et la Formation)*, ATIEF, 2011, 18, 14 p. hal-00696375
- [22] Zinn, C.: Algorithmic debugging to support cognitive diagnosis in tutoring systems. In: *Proceedings KI 2011: Advances in Artificial Intelligence*. LNCS, vol. 7006, pp. 357368 (2011)
- [23] Self, J. (1999). The defining characteristics of intelligent tutoring systems research: ITSs care, precisely. *International journal of artificial intelligence in education*, 10(3-4), 350-364.
- [24] Bourdeau, J., & Grandbastien, M. (2010). Modeling tutoring knowledge. In *Advances in intelligent tutoring systems* (pp. 123-143). Springer, Berlin, Heidelberg.
- [25] Pelleu-Tchétagani, J. M. (2005). Une approche propéagogique du diagnostic cognitif dans les STI: conception, formalisation et implémentation.
- [26] Roger Nkambou, Jacqueline Bourdeau, and Riichiro Mizoguchi (Eds), "Advances in Intelligent Tutoring Systems", 2010.
- [27] Almasri, A., Ahmed, A., Almasri, N., Abu Sultan, Y. S., Mahmoud, A. Y., Zaqout, I. S., ... & Abu-Naser, S. S. (2019). *Intelligent tutoring systems survey for the period 2000-2018*.
- [28] Abdelrahman, G., Wang, Q., & Nunes, B. P. (2022). Knowledge Tracing: A Survey. arXiv preprint arXiv:2201.06953

# Improving the Diabetes Diagnosis Prediction Rate Using Data Preprocessing, Data Augmentation and Recursive Feature Elimination Method

E. Sabitha, M. Durgadevi

Dept. of Computer Science and Engineering, College of Engineering and Technology  
SRM Institute of Science and Technology Vadapalani, Campus No.1, Jawaharlal Nehru Road, Vadapalani TN, India

**Abstract**—Hyperglycemia is a symptom of diabetes mellitus, a metabolic condition brought on by the body's inability to produce enough insulin and respond to it. Diabetes can damage body organs if it is not adequately managed or detected in a timely manner. Many years of research into diabetes diagnosis has led to a suitable method for diabetes prediction. However, there is still scope for improvement regarding precision. The paper's primary objective is to emphasize the value of data preprocessing, feature selection, and data augmentation in disease prediction. Techniques for data preprocessing, feature selection, and data augmentation can assist classification algorithms function more effectively in the diagnosis and prediction of diabetes. A proposed method is employed for diabetes diagnosis and prediction using the PIMA Indian dataset. A systematic framework for conducting a comparison analysis based on the effectiveness of a three-category categorization model is provided in this study. The first category compares the model's performance with and without data preprocessing. The second category compares the performance of five alternative algorithms employing the Recursive Feature Elimination (RFE) feature selection method. Data augmentation is the third category; data augmentation is done with SMOTE Oversampling, and comparisons are made with and without SMOTE Oversampling. On the PIMA Indian Diabetes dataset, studies showed that data preprocessing, RFE with Random Forest Regression feature selection, and SMOTE Oversampling augmentation can produce accuracy scores of 81.25% with RF, 81.16 with DT, and 82.5% with SVC. From Six Classifiers LR, RF, DT, SVC, GNB and KNN, it is observed that RF, DT, and SVC performed better in accuracy level. The comparative study enables us to comprehend the value of data preprocessing, feature selection, and data augmentation in the disease prediction process as well as how they affect performance.

**Keywords**—Artificial Intelligence (AI); Machine Learning (ML); Deep Learning (DL); Neural Network; Diabetes Mellitus; Recursive Feature Elimination (RFE); Synthetic Minority Over-sampling Technique (SMOTE)

## I. INTRODUCTION

A metabolic disorder known as diabetes mellitus is characterized by hyperglycemia brought on by the body's inability to create and utilize insulin.[1]. There are three forms of diabetes types. The human body cannot generate enough insulin when it has type I. The body is unable to produce or use insulin effectively in type II. During pregnancy, gestational diabetes can develop [2]. Both Type I and Type II diabetes are getting more and more prevalent worldwide, with

Type II diabetes being at epidemic levels. According to medical study, diabetes has been linked to the long-term degradation of vital organs. More concerning is its impact on pregnancies: diabetes affects roughly 7% of pregnancies each year, posing a dual life-threatening risk. Over half of the world's population is expected to have diabetes by 2045 due to the disease's rising prevalence. The WHO predicted that 463 million people will be diabetic worldwide by 2020, and these are just the cases that have been identified. In United States almost one in every 10 individuals is diabetic. Diabetes research is therefore essential, including studies of diabetes prediction and its effects on health [3].

Diabetes can be diagnosed by either an oral glucose tolerance test result or a fasting plasma glucose level. On the other hand, diabetes can be identified by Glycemic threshold levels. This is due to the fact that different ethnic groups have varied risk levels. Multiple blood sugar tests are taken both before and after a meal. By observing a relevant decision at a time, practitioners are faced with the difficult task for diagnosing diabetes. The diagnostic process, on the other hand, can be made more computationally simple [4]. The fields of technology and medicine have been profoundly affected by big data and data analytics approaches. Rather than depending on conventional methodologies, which are usually unable to handle massive data, cutting-edge technologies like ML, DL and cloud computing must be employed to fully utilize the data and automate computation processes in medical research. This paper provides a customised hybrid model of artificial neural networks (ANN) and genetic algorithms as a framework for accurately forecasting the onset of diabetes, replete with regularisation and prediction techniques created for diabetes prediction [5].

Numerous computational projects have been started recently, many of which are focused on the use of ML and DL algorithms in diabetes research with the goal of assisting physicians in making rapid and accurate diagnosis decisions. With the ongoing development of diabetes testing equipment, individuals can now take part in individualised examinations of their diabetes status for better lifestyle modifications. In comparison to existing methods, a dependable accuracy rate is categorized in recent studies. A higher accuracy rate in diabetes prediction is essential, as early diagnosis of diabetes mellitus is required. The researchers are presenting a range of DL and ML methods for diabetes forecasting. Despite a large amount of research on diabetic prediction, the accuracy still

needs to be improved. This is necessary since diabetes poses major health risks if it is not effectively treated or diagnosed in a timely manner. In this paper a comparative evaluation is done based on feature selection approaches and data augmentation techniques that increase prediction performance in this research. The main contribution of the paper is summarised as follows:

- 1) The significance of data pre-processing is demonstrated by comparing the outcomes of the proposed model with and without data pre-processing.
- 2) To emphasise the significance of feature selection in disease prediction, which improves model performance and boosts predictive power.
- 3) To overcome the issue of a small dataset, data augmentation is employed to enhance the dataset's size. Deep learning and machine learning typically require a large quantity of data to train the networks.

## II. RELATED WORK

The views of data pre-processing, data augmentation, and classification are the foundation of the current body of work. However, in this paper the review is limited to the recent publications. Diabetic research has recently begun to improve based on the performance accuracy. This article can be used by readers to learn about the past and present effectiveness of algorithms in diabetes research [6]. Table I illustrate the review on recent papers that work on the diabetic prediction. The review is based on the recent trends and papers that are suitable for the disease prediction. In diabetes research, the NN-based approaches have constant to increase accuracy. The problems of data standardisation, imbalance, and feature augmentation are addressed in this Min-Max Normalisation and a Variational Autoencoder [7]. MLP was then utilised for classification, with an accuracy rate of 92.31 percent. In [8], the accuracy of their Artificial Backpropagation Scaled Conjugate Gradient Neural Network (ABP-SCGNN), which was previously reported to attain 93 percent accuracy without data pre-processing, has significantly improved. The work of [9] demonstrates another impressive result with NN-based models. They looked at iterative imputers, k-nearest neighbour (K-NN), and median value imputation. In order to acquire an F1-score of 98 percent, MLP was then employed for classification. For feature selection and missing value imputation in [10], Pearson correlation and median value imputation were used. The authors used interquartile ranges to further normalise the data and eliminate outliers. DNN-based classification model, which contained a number of hidden layers, was 88.6% accurate. A deep neural network (DNN) model's accuracy was estimated to be 98.07 percent in [11]. Even though the authors claim to have used data cleansing, the process is not described in the article. In [12] used the median value for missing value imputation and principal component

analysis (PCA) for feature selection. MLP was then used to carry out the classification procedure, and it had a 75.7 percent accuracy rate. For feature selection and missing value imputation, PCA and minimum redundancy, maximum relevance (mRMR) was also used in [13]. Using an MLP, they were able to get a classification accuracy of 73.90%. Many different methods were employed [14]; implemented many assessment techniques, including Nave Bayesian, Random Forest (RF), KNN, and K-fold Cross-Validation. The technique has a 64.47 percent accuracy, according to K-fold Cross Validation. Nave Bayes, function-based multilayer perceptrons, and RF based on decision trees were all employed in [15]. The feature extraction method was utilised to extract reliable and illuminating properties from the dataset using the correlation method. According to the author, the Nave Bayes method outperformed the random forest and multilayer perceptron algorithms.

In [16] tested various machine learning methods for predicting early diabetes on the PID dataset. Using 20-fold cross-validation and a 70-30 train-test split, tree-based RF scored 75.65 percent, Nave Bayes (NB) 71.74 percent, and KNN 65.19 percent. A decision tree and the gradient boosting method were used by [17] for prediction. The technique has a classification accuracy of 90% and computes a correlation value to determine differences between a diabetic patient and healthy person. With 10-fold-cross validation and an enhanced K-Means cluster method, [18] obtained 95.42 percent accuracy. In [19] used 10-fold cross validation with machine learning techniques on patients who had a history of non-diabetics and a cardiac problem. In [20], which used machine learning as a prediction model for type 2 diabetes mellitus early prediction, Glnet, RF, XGBoost, and Light all shown improved clinical prediction. It is suitable for one dataset but inappropriate for another. A more advanced DNN-based diabetes risk prediction model that not only predicts but also identifies who will develop the ailment in the future was proposed in [21]. Before training on several classification models, such as NB, LR, RF, AB, GBM, and extreme gradient boosting, the mean of each column of data was pre-processed in [22] to remove missing values. With a precision of 77.54 percent, the XGBoost model was the most precise. The efficiency of the classification models SVM, K-NN, NB, Gradient boosting (GB), and RF were contrasted in [23]. With an accuracy of 98.48 percent, the RF prevailed. In [24] employed Pearson correlation for feature selection and mean value imputation for missing value. The authors assessed the performance of various classification models, including extreme boosting (XB), AB, RF, DT, and K-NN, using a K-fold cross-validation environment and the grid search strategy for hyperparameter tuning. With an accuracy percentage of 94.6 percent, the XB won. Linear SVM, Radial Basis function SVM, DT, and K-NN were employed in a stacked ensemble to achieve a classification accuracy of 83.8%.

TABLE I. OVERVIEW OF THE LITERATURE REVIEW

Authors	Feature selection (FS)	Classification	Comments
Benavides, C., et al .[7]	FS: none specified;removed missing values;	MultiLayer Perceptron	MLP achieved the best accuracy, 92.31%
Alkhamees, B. F et al.[8]	FS: none specified; MVI: none specified	ANN trained with ABS conjugate gradient neural network (ABP-CGNN)	Achieved 93% accuracy
Ahmad, M., et al.[9]	Median value, K-NN, and iterative imputer were used for missing value imputation	ANN	ANN achieved 98% accuracy
Foo, S. Y et al.[10]	FS: Pearson correlation MVI: Median value for missing values imputation.	DNN run with different hidden layers	Achieved 86.26% accuracy with 2 hidden layers.
Naz, H., & Ahuja, S. [11]	Method not stated	MLP and DL with 2 hidden layers	DL achieved best accuracy of 98.07%
Iqbal, M. A., [12]	FS: PCA; MVI: Median value	MLP	Achieved 75.7% accuracy
Qu, K., et al. [13]	FS: PCA; MVI: redundancy and minimum relevance	MLP	Achieved 73.90% accuracy
Halgamuge, M. N., et al.[14]	none specified	NB,RF,KNN,K-fold cross validation	Using K-fold CrossValidation, the method achieved 64.47% accuracy.
Singh, D. A. A. G., et.al.[15]	Correlation method	decision tree-based RF, function-based multilayer perceptron ,Naïve Bayes	Naïve Bayes algorithm achieved better results
Awais, M., et.al.[16]	none specified	RF,NB,KNN with 20 -fold cross-validation	RF achieved 75.65%
Selvan, K. A., et.al.[17]	none specified	DT,GB	Achieved 90% accuracy
Yang, S., et.al.[18]	none specified	K-Means cluster algorithm with 10 fold-cross validation.	Acheived 95.42% accuracy
Gnanadass [22]	none specified	NB, linear regression (LR), RF, AB, gradient boosting machine (GBM), and extreme gradient boosting (XGBoost).	XGBoost achieved 77.54%
Mounika, B., et al.[23]	none specified	SVM, K-NN, NB, GB, RF, LR	RF achieved best accuracy of 98.48%
Hasan et al [24].	FS: correlation; MVI: mean value	XB, AB, RF, DT, K-NN	XB achieved best accuracy of 94.6%

### III. MATERIALS AND METHODOLOGY

#### A. Dataset

The PIMA Indian Diabetes database was used for this study. The main objective of the dataset is to establish a patient's diagnostic diabetes status. The dataset contains one outcome variable and a number of medical predictor variables. Predictor variables for diabetes include age, number of pregnancies, BMI, BP, glucose, Skin thickness, Insulin, and Diabetes pedigree function. Particularly, all of the patients are females in PIMA who are at least 21 years old. The selection of these examples from a broader database was subject to several of limitations. Our proposed research compares data pre-processing, feature selection, and data augmentation techniques. The study aims to overcome flaws in early diabetes mellitus diagnosis that impair accuracy. The disadvantages are as follows: 1. A large number of missing values lead to erroneous predictions. 2. Imbalanced data has an impact on the model's performance. The suggested framework illustrates each stage of prediction work, including data pre-processing, Feature selection techniques incorporated with the Recursive Feature Elimination approach, and Smote

data augmentation with and without Smote data. The study compares model accuracy by applying the augmentation strategy to improve the dataset. The study compares not only on the basis of augmentation, but also on the basis of feature selection strategies. Fig. 1 shows a diagrammatic representation of the planned work.

#### IV. PROPOSED WORK

In this paper, a systematic framework is proposed for the diabetic prediction with different classifiers. The importance of the data preprocessing is elaborated by presenting the results obtained. The main contribution of the proposed work is to show the importance of data cleaning by using the data preprocessing strategies then by selecting the important attributes that highly correlate with diabetic prediction. Most important part is balancing the dataset by SMOTE data augmentation, the proposed work focus on the three part that improves the prediction accuracy. The six classifier algorithm are then used for classification. In the proposed system, inconsistent data is replaced, then suitable features are selected by using the RFE with the Random Forest Regression and finally the selected are augmented by SMOTE



Oversampling technique to improve the imbalance dataset problem.

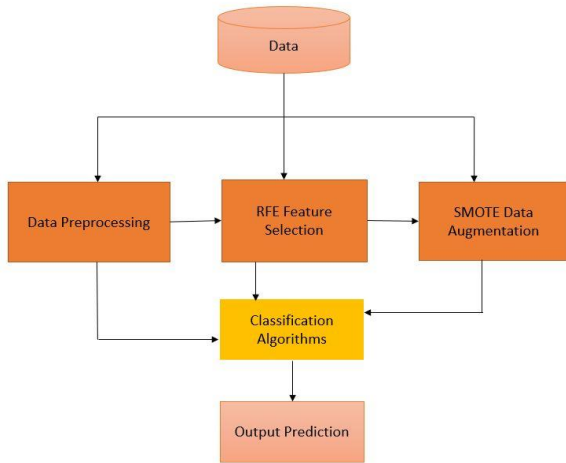


Fig. 1. Proposed Work.

### V. DATA PRE-PROCESSING

The idea of data pre-processing refers to the conversion of unclean data into a clean data set. The dataset is pre-processed to look for missing values, noisy data, and other irregularities before applying the algorithm. These data are crucial for decision-making, thus accurate and effective estimate techniques are required. For this analysis, PIMA Indians Diabetes database is taken. The dataset has more missing values than null values. In the medical field, the problem of a database with missing values is very widespread. The Table II displays the number of zeros in each attribute, while Fig. 2 and Fig. 3 illustrate the proportion of missing data and the impossible value assigned in the pregnant feature, both of which reduce the model's performance. Using data pre-processing techniques, the study focuses on cleaning up the data by improving the values assigned to each feature. In this paper no specific strategies are followed to clean up the data in the suggested work, instead a few simple and easy ways to pre-processing is done to clean up and improve the quality of the data. The following are the methods that will be described.

TABLE II. MISSING VALUE IN DATASET

Features	Total	Percent
Insulin	374	48.697917
Skin Thickness	227	29.557292
Blood Pressure	35	4.557292
BMI	11	1.432292
Glucose	20	0.651042
Pregnancies	0	0.00
Diabetes Pedigree Function	0	0.00
Age	0	0.00
Outcome	0	0.00

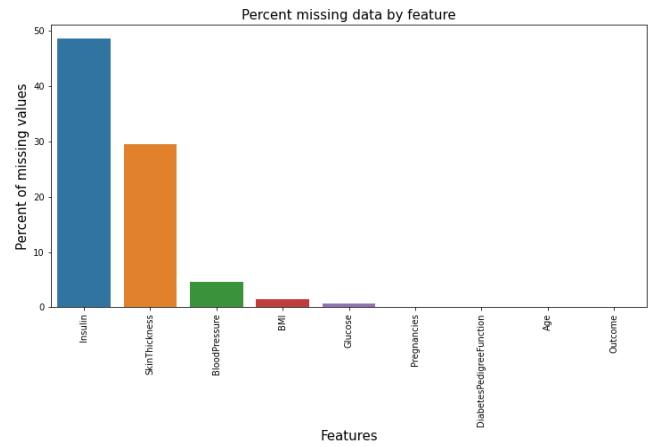


Fig. 2. Percent of Missing Data in Features.

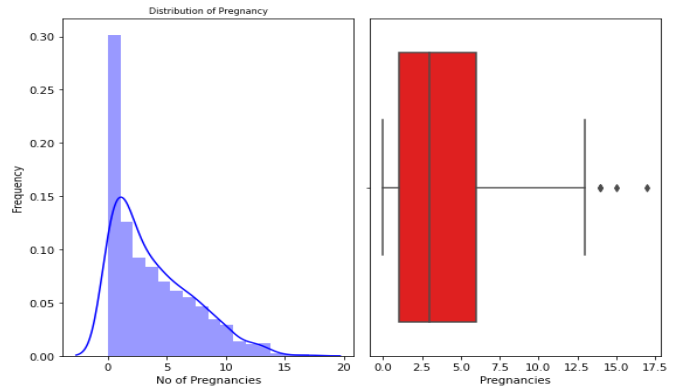


Fig. 3. Outliers of Pregnancies Feature.

#### A. Treatment of Missing Values

If the behaviour and linkages with other variables are not adequately analysed, missing data in a data collection could reduce a model's power or fit or result in a biased model. The classification or prediction that results could be inaccurate. When the dataset's above mentioned attributes were evaluated, it was found that several of them had zero values and that the pregnancy variable had a maximum value of 17, which seemed to be impossibly high. There is a range for a typical healthy human being that is not zero, suggesting a missing value, hence these 0 column values are illogical. To make counting the missing integers simpler, we'll start by swapping out these zeros for NaN. Later, we'll swap them out for the proper values. There are a number different ways to handle missing values. The choices are displayed below. Simply deleting all instances where the variable being considered contains missing values is the simplest method to handle with missing values. However, this approach can mean losing actually valued data about patients. The calculation of mean is the second approach to complete all gaps in the data. In this approach the missing values are replaced by with the average value calculated in the same attribute. This method can reduce the loss of data instead of removing the missing values that reduces the quality of the dataset. In this method all the missing values are replaced by zero. The process of replacing by zero is simply by replacing any missing values with zero. Since the data in this study have been converted to values

between zero and one, substituting zero for missing values has the same result as substituting the attribute's lowest value. However this method leads to poor classification by missing data which are inaccurately appraised if they are necessary for clinical management. The K-nearest neighbour method is the fourth method which is used to replace the missing values in the dataset. The missing values are replaced with the value of nearest-neighbor column. The nearest-neighbor column is considered to be the closest column in Euclidean distance. The next closest column is used if the relevant value from the nearest-neighbor column also contains a missing value.

The Table I clearly describes the PIMA dataset. The features like Glucose, Blood Pressure, Skin thickness, Insulin, and BMI are with 0 values. In this study, we use second approach i.e. all missing values of an attribute are replaced by the mean by calculating the average of all accessible values of the same attribute. When all the zero values are replaced with mean value, the dataset was further split into training and testing data. The dataset as a whole is made up of 80% training data and 20% test data. The model performance is evaluated by the model accuracy, which is determined via machine learning algorithms. On two levels—one where the zero values were replaced with the mean and another where they weren't—we compared the model's performance. By contrasting the two, we can see how useful data pre-processing is in improving the dataset's suitability for subsequent operations. The comparison can be seen in upcoming session.

### B. Without Data Pre-Processing

The PIMA dataset is directly used in the machine learning algorithm to assess the prediction's accuracy without any data pre-processing. Some of the techniques used are Gaussian Nave Bayes (GNB), KNN, DT, Support Vector Classification (SVC), LR, and RF. The dataset is used in the classification approach to determine the degree of accuracy in disease prediction that may be made without any prior processing. Table III and Fig. 4, clearly represent the performance of the classification algorithm based on the accuracy. The accuracy for LR was 77.5 percent with 0.06 training time, for RF it was 76.5 percent with 0.77 training time, for DT it was 67.5 percent with 0 training time, for SVC it was 81.2 percent with 0.03 training time, and for GNB and KNN it was 75% with 0.02 training time.

TABLE III. ACCURACY OBTAINED WITHOUT DATA PRE-PROCESSING

S.no	Classifier	Validation accuracy	Training time
0	Logistic Regression	0.775	0.06
1	Random Forest	0.7625	0.77
2	Decision Tree	0.675	0
3	SVC	0.8125	0.03
4	GaussianNB	0.75	0.02
5	KNeighbors	0.75	0.02

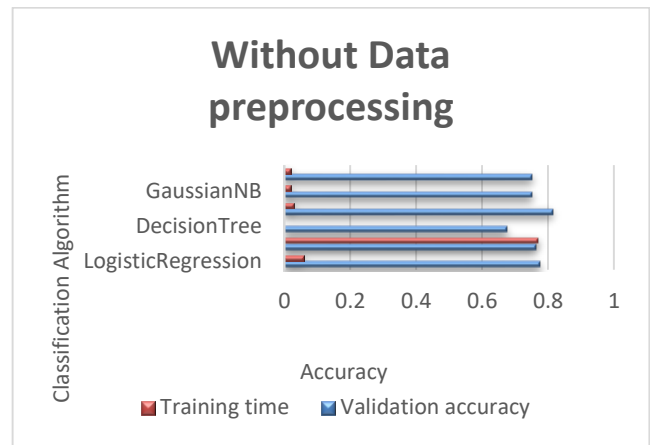


Fig. 4. Accuracy Without Data Pre-processing

### C. With Data Pre-processing

1) *Dropout missing values*: The simple method to deal with missing values, is to simply delete any instances in which the variable being studied contains missing values. The loss of potentially relevant data regarding patients whose values are missing, however, could be a consequence of this strategy. The dataset is divided in an 80:20 ratio between training and testing data, after all missing values have been removed, and the features are chosen using RFE with Random Forest regression. An imbalance dataset generally speaking is the PIMA dataset. We can see that 268 people have diabetes and 500 people do not when we analyse the dataset by outcome.

The performance of the classification algorithm worsens when the data for training and testing are split due to the unequal size of the training and testing sets. Using classification techniques to determine correctness, we augment the dataset with additional data to address the imbalance issues. Table IV and Fig. 5 represent the result obtained. The accuracy scores are LR 69.04 percent with 0.05 training time, RF 85.7 percent with 0.62 training time, DT 73.8 percent with 0 training time, SVC 76.1 percent with 0.01 training time, GNB 73.8 percent with 0.02 training time, and KNN 71.42 percent with 0.01. When we focus on accuracy, RF has achieved the maximum score of 85.7 percent, but the training duration is 0.62 minutes. When it comes to training time, DT, SVC, GNB, and KNN take less time, but their accuracy is worse than RF.

TABLE IV. ACCURACY OBTAINED WITH DROP OUT MISSING VALUES

S.no	Classifier	Validation accuracy	Training time
0	Logistic Regression	0.690476	0.05
1	Random Forest	0.857143	0.62
2	Decision Tree	0.738095	0
3	SVC	0.761905	0.01
4	GaussianNB	0.738095	0.02
5	KNeighbors	0.714286	0.01

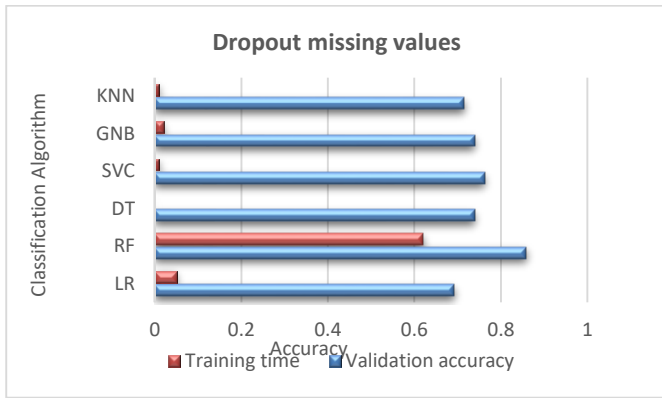


Fig. 5. Accuracy with Dropout Missing Values.

2) *Replacing missing value with mean*: There are no null values in the PIMA dataset, however there are more missing values, such as zero values in many characteristics, as previously indicated. As a result, the model's performance may be affected. To solve this problem, one option is to remove the zero values, but this reduces the algorithm's performance. Instead, in this study, we can replace the zero by calculating the attribute mean values and replacing the zero. One way to pre-process a dataset without reducing its size is by using this technique. Following pre-processing, the dataset is divided into train and test groups in an 80:20 ratio. The next stage is to assess the accuracy of the diabetic prediction using classification algorithms.

TABLE V. ACCURACY OBTAINED WITH MEAN VALUE

S.no	Classifier	Validation accuracy	Training time
0	Logistic Regression	0.725	0.04
1	Random Forest	0.7625	0.52
2	Decision Tree	0.7	0
3	SVC	0.7625	0.02
4	GaussianNB	0.7	0.01
5	KNeighbors	0.7625	0.01

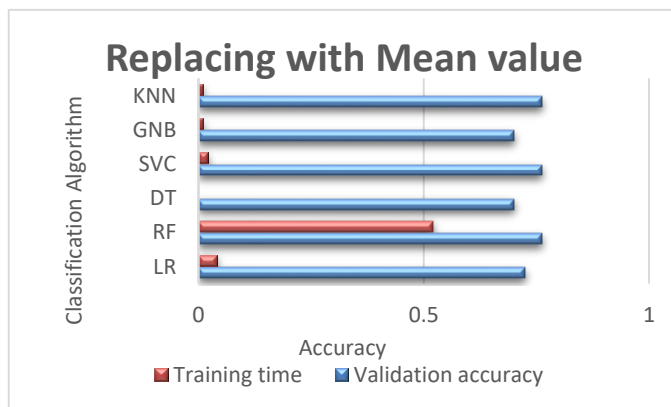


Fig. 6. Accuracy with Replacing Mean Value.

The accuracy of classification algorithms when the inconsistent value is replaced by Mean value is explained in Table V and graphically presented in Fig. 6. When the dataset is pre-processed, we can witness a gradual improvement in accuracy in Tables IV and V. The dataset was pre-processed in this study by removing zero values and replacing them with the mean of the characteristics. There isn't much of a change, however pre-processing the dataset can help us gain a little more precision. Understanding the significance of data pre-processing in any research effort is made easier by this approach. Data pre-processing is essential for evaluating the model's performance or determining the algorithm's efficiency.

## VI. FEATURE SELECTION

Finding the most useful set of features for creating efficient models of the phenomenon being studied is the goal of feature selection. Feature selection strategies are divided into two categories: i) supervised techniques and ii) unsupervised procedures. The efficiency of supervised models is increased by using labelled data in supervised procedures to find pertinent characteristics. Unlabelled data is utilised in unsupervised approaches. In terms of taxonomic classification, these methods fall under the headings of A) Filter methods, B) Wrapper methods, C) Embedded methods, and D) Hybrid methods.

### A. Recursive Feature Elimination

To choose the features in the study, a Wrapper method based Recursive feature elimination (RFE) strategy is applied. RFE is a greedy optimization strategy that selects features by considering a reduced set of features iteratively. A variety of deep learning algorithms are provided and employed in the method's core to choose features. On the other hand, filter-based feature selections rank each feature according to its importance and choose the ones with the highest or lowest scores. The given algorithms, such as random forest, decision trees, and SVM, are used to score features, or a more general technique that is independent of the whole model is used. The importance of the features used in training the estimator is decided using the feature importance attribute. Until we get the required number of features, the least important feature is deleted from the existing collection of features.

#### Procedure

Step 1: Fit the RFE method to the model

Step 2: The feature importance attribute is used to rank features.

Step 3: Once the necessary number of features is collected, the least significant feature is deleted and the procedure is repeated.

#### RFE with five different algorithms

In this study feature selection done five different ways.

- 1) Manual feature selection
- 2) RFE with Logistic regression
- 3) RFE with Random Forest regression
- 4) RFE with Decision Tree regression

- 5) RFE with Decision Tree Classifier
- 6) RFE with 5-Cross validation

### B. Manual Feature Selection

The features for the prediction are manually picked in manual feature selection. The outcome is one of nine attributes in PIMA, eight of which are independent. We chose six attributes out of the eight for this research. The characteristics were chosen after reviewing a large number of research articles that were more accurate in diabetic prediction. We choose six attributes manually like Pregnancy, Glucose, Blood Pressure, Insulin, BMI, and Age. The dataset is then split in half, 80:20, into train and test sets. The classification algorithms such as GNB, KNN, DT, SVM, and RF are utilized, to analysis the performance quantified in terms of accuracy. The dataset, methodology, and accuracy achieved are all evaluated in the Table VI (also see Fig. 7).

### C. Recursive Feature Elimination with different Methods

In this research, RFE was used in conjunction with several methods such as logistic regression. The key characteristics from the dataset are selected using Random Forest regression, Decision tree regression, Decision tree classifier, and RFE with five-fold cross validation as an estimator. Pregnancies, BMI, DPF, and Glucose are the most common features selected by each feature selection technique. Pregnancies, BMI, DPF, and Glucose are four of the eight variables that are thought to be directly connected to diabetic prediction. The dataset is then divided into a train set and a test set in an 80:20 ratio based on the selected attributes. Classification algorithms like GNB, KNN, DT, SVM, and RF are used for prediction, and the performance of the algorithm is evaluated in terms of accuracy. The dataset, feature selection procedure, selected features, Classification Algorithm, and Accuracy are all detailed in Table VII

TABLE VI. MANUAL FEATURE SELECTION

Dataset	Classification Algorithm	Accuracy
PIMA	GNB	77.27%
	KNN	75.97%
	DT	70.10%
	SVM	79.87%
	RF	81.81%

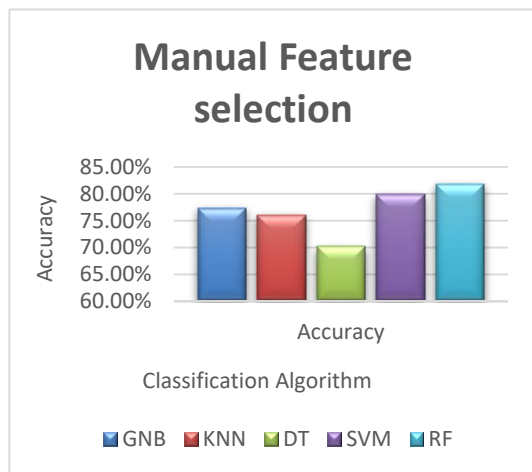


Fig. 7. Accuracy with Manual Feature Selection.

RFE with Logistic Regression selected Pregnancies, BMI, and DPF features after observing several feature selection methods. The features are then divided into two groups: training and testing. For classification and accuracy evaluation, algorithms such as GNB, KNN, DT, SVM, and RF are utilized. GNB obtained 79.87 percent accuracy using the algorithms. RFE with Random Forest Regression, using Glucose, BMI, and DPF as selected features, has a greater accuracy of 81.16 percent when compared to DT. The same features were chosen from RFE with Decision Tree Regression and RFE with Decision Tree Classifier. KNN achieved 75.32 percent accuracy by using both selection methods. Pregnancies, Glucose, and BMI were chosen as characteristics for RFE with five cross validations. The KNN algorithm achieved 79.2 percent accuracy based on the features. When compared to other algorithms, RFE with Random Forest Regression utilizing DT has obtained greater accuracy of up to 81 percent, according to the detailed analysis using the RFE feature selection method and algorithm.

## VII. DATA AUGMENTATION

Data augmentation is a series of techniques for generating extra data points from existing data in order to fictionally increase the amount of data accessible. Simple data modifications or the use of deep learning models to generate more data are instances of this. Applications for machine learning are quickly increasing and diversifying, especially in the deep learning space. Approaches for data augmentation may be effective in the struggle against the drawbacks of artificial intelligence. One step in building a data model is cleaning the data, which is necessary for high accuracy models. The model won't be able to produce reliable predictions for inputs from the real world, though, if data cleansing limits representability. In order to increase the reliability of machine learning models, data augmentation techniques can be employed to replicate variations that the models would encounter in the actual world.

### A. SMOTE Oversampling

In many disciplines, unbalanced data has been a problem, causing most approaches to produce erroneous forecasts that strongly favour the dominant class. To decrease the harmful impact of unbalanced data, we can optimise the process using a variety of techniques: Certain techniques, like as oversampling, under sampling, or both, are employed to correct the unbalanced data set in order to generate a balanced distribution. A statistical method for equally expanding the number of cases in a dataset is called SMOTE (Synthetic Minority Oversampling Technique). Based on current minority conditions, the component creates new instances. The overfitting issue brought on by random oversampling is helped by the SMOTE algorithm. The working approach begins by setting up the total number of oversampling observations N. A binary class distribution of 1:1 is typically used to select it. This could be minimised, though, depending on the circumstances. After that, a positive class instance is randomly chosen and the loop starts. Then, the KNNs for that instance are obtained. In order to interpolate new synthetic



instances in the end, N of these K instances are chosen (see Fig. 8).

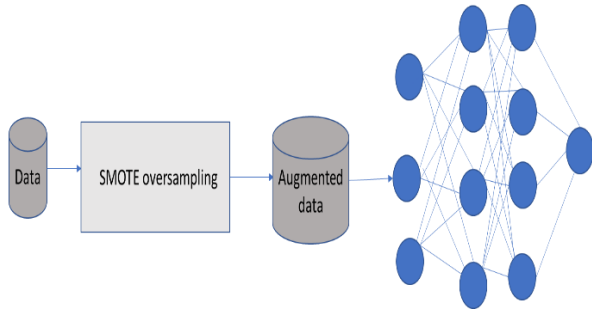


Fig. 8. SMOTE Oversampling.

**Algorithm**

Step 1: To interpolate new synthetic instances in the end, N of these K instances are chosen.  $x \in A$ .

Step 2: The uneven proportion determines the sampling rate N. N samples ( $x_1, x_2, \dots, x_n$ ) are chosen at random from  $x \in A$ . Set A1 is composed of the k-nearest neighbours.

Step 3: For each  $x_k$ , a new example is produced using A1 ( $k = 1, 2, 3, \dots, N$ ):  $x^{\wedge} = x + \text{rand}(0,1) * |x - x_k|$  in which  $\text{rand}(0,1)$  represents the random numbers between 0 and 1.

**B. SMOTE Oversampling with RFE with Random Forest Regression**

In the PIMA dataset, the result is in an unbalanced state. When examining the outcome, 1 counts to 268 and 0 counts to

500, resulting in false in excess of true. The model is trained with a higher percentage of false values than true values. SMOTE oversampling is used to reduce the data's complexity and balance it. The features selected through the RFE with Random Forest regression feature selection technique are subjected to oversampling. The Random Forest algorithm includes a feature importance calculation that can be done in two ways. The Gini coefficient is calculated using the Random Forest structure. Decision Tree algorithm with internal nodes and leaves make up each decision tree that makes up a Random Forest. The internal node uses the chosen characteristic to determine how to divide the data set into two sets with similar replies. For classification tasks, criteria like gini impurity or information gain, as well as variance reduction for regression, are used to select the internal node properties. The importance of a feature is determined by the average of all trees in the forest. There is also Mean Decrease. Accuracy is a method for calculating the importance of features on permuted out of bag samples based on the accuracy's mean reduction. The scikit-learn package does not include this function. The selected features from the RFE using Random Forest Regression are Glucose, BMI, and DiabetesPedigreeFunction. The features are then enhanced based on their results, and performance is measured using machine learning algorithms such as Logistic Regression (LR), RandomForest(RF), DecisionTree(DT), SVC, GaussianNB, and KNeighbor's for further classification. The accuracy attained with and without data augmentation is shown in the table VIII and IX.

TABLE VII. RFE WITH DIFFERENT FEATURE SELECTION METHOD

Dataset	Feature selection method	Selected Features	Classification Algorithm	Accuracy
PIMA	RFE with Logistic Regression	Pregnancies BMI DPF	GaussianNB(GNB) Knearestneighbor(KNN) Decision Tree(DT) SupportVectorMachine(SVM) RandomForest(RF)	GNB: 79.87 % KNN: 72.72 % DT: 62.98 % SVM: 72.07 % RF: 71.42%
	RFE with Random Forest Regression	Glucose BMI DPF	GaussianNB(GNB) Knearestneighbor(KNN) Decision Tree(DT) SupportVectorMachine(SVM) RandomForest(RF)	GNB : 74.02 % KNN : 72.77 % DT : 81.16 % SVM : 75.32 % RF : 75.97%
	RFE with Decision tree Regression	Glucose BMI DPF	GaussianNB(GNB) Knearestneighbor(KNN) Decision Tree(DT) SupportVectorMachine(SVM) RandomForest(RF)	GNB : 74.02 % KNN : 75.32 % DT : 68.27 % SVM : 74.67% RF : 72.72%
	RFE with Decision tree Classifier	Glucose BMI DPF	GaussianNB(GNB) Knearestneighbor(KNN) Decision Tree(DT) SupportVectorMachine(SVM) RandomForest(RF)	GNB : 74.02 % KNN : 75.32 % DT : 68.27 % SVM : 74.67% RF : 72.72%
	RFE with five cross validations	Pregnancies Glucose BMI	GaussianNB(GNB) Knearestneighbor(KNN) Decision Tree(DT) SupportVectorMachine(SVM) RandomForest(RF)	GNB : 74.02 % KNN : 79.2 % DT : 70.7 % SVM : 75.9% RF : 72.72%

TABLE VIII. WITH SMOTE AUGMENTATION

S.no	Classifier	Validation accuracy	Training time
0	Logistic Regression	0.80	0.03
1	Random Forest	0.7875	0.72
2	Decision Tree	0.725	0
3	SVC	0.825	0.03
4	GaussianNB	0.775	0.02
5	KNeighbors	0.7125	0.02

TABLE IX. WITHOUT SMOTE AUGMENTATION

S.no	Classifier	Validation accuracy	Training time
0	Logistic Regression	0.69	0.05
1	Random Forest	0.70	0.59
2	Decision Tree	0.72	0
3	SVC	0.7096	0.02
4	GaussianNB	0.6774	0.02
5	KNeighbors	0.6935	0.02

We can compare the accuracy gained with and without augmentation using the two tables above. The accuracy and training duration are shown in Table VIII with moderate improvement. The accuracy and training time are displayed in Table IX without any augmentation. When we compare the two tables, it's evident that by supplementing the data, we can improve accuracy. When compared to various machine learning algorithms, SVC with smote augmentation had the highest accuracy of 82.5 percent with a training time of 0.03. Whereas Logistic regression scored 80 percent with a training time of 0.03, Random Forest scored 78.75 percent with a training time of 0.72, Decision Tree scored 72.5 percent with a training time of 0, GaussianNB scored 77.5 percent with a training time of 0.02, and Kneighbors scored 71.25% with a training time of 0.02.

### VIII. RESULTS

The work is primarily focused on demonstrating the significance of data preprocessing, feature selection, and data augmentation in disease prediction that have a significant impact on the model's performance. Six distinct classification algorithms are used in the comparison analysis to highlight the impact with and without Data pre-processing, SMOTE Oversampling.

The work is divided into three sections:

- 1) With and without data pre-processing, and utilising classification methods to assess accuracy.
- 2) For feature selection, combining RFE with logistic regression, Random Forest regression, RFECV, Decision tree regression, and Decision tree classifier
- 3) Using SMOTE oversampling to improve accuracy by decreasing concerns caused by dataset imbalance.

In order to prepare the PIMA and diabetes type data for a deep learning model, data pre-processing is required. The PIMA dataset contains missing values, which means that many attributes have fewer values, such as zeros. The erroneous values reduce the model's performance. The number of null values in each attribute is shown in Table II. Properties including blood pressure, skin thinning, glucose, insulin, and BMI are all zero in Table II, along with other characteristics. Two data pre-processing strategies are used to replace zero values: one removes the zero values and the other replaces the values by finding the mean. Table III shows the results obtained without pre-processing the data. Tables IV and V show the outcomes of two different approaches to data pre-processing. We employ six classifiers in this work for data pre-processing: LR,GNB,KNN,RF,DT, and SVC. When comparing the results, we can see that when the data set is pre-processed, RF achieves an accuracy of 85.7 percent. Several algorithms, including LR, Random Forest Regression, Decision Tree Classifier, Decision Tree Regression, and RFE with cross validation, were used in our work to use the RFE feature selection approach. Following feature selection, the accuracy is determined using five classifiers. When we compare the results of the classifier based on feature selection RFE with Random Forest regression, we find that Random Forest regression has a better outcome DT: 81.16 percent. The third comparison is based on data augmentation with SMOTE Oversampling versus data augmentation without SMOTE Oversampling. The data augmentation method is used to alleviate the issue caused by an unbalanced dataset. Following feature selection, the SMOTE Oversampling technique is used to augment the selected feature dataset in our study. After that, six classifiers are used, and the accuracy of the classifiers is measured. When we examine the results of both methods, with and without SMOTE oversampling, we can see that with SMOTE oversampling, we can attain higher accuracy. Table VIII and Table IX explains the results obtained with and without SMOTE oversampling.

### IX. CONCLUSION

The comparison is based on three categories, which are elaborated in the study: (i) with and without data pre-processing, (ii) feature selection using five alternative algorithms, (iii) with and without data augmentation. The importance of data pre-processing, feature selection, and data augmentation can be seen in the three comparisons. When each category is examined separately, data pre-processing comes out on top since it significantly affects how well models or algorithms perform. When we pre-process a dataset, the dataset's quality improves, as does the performance of the models or algorithms. Tables III, IV, and V provide a comprehensive explanation of the contrast. In this study, pre-processing is done in two different ways, and the results of the two methods are compared. We used five different algorithms with RFE and compared the results in Table VII. Likely Feature Selection Importance was also thoroughly explained, and we used five different methods with RFE and compared the results in Table VII. When it comes to data augmentation, the goal is to solve problems that arise from an unbalanced dataset. The work utilised the SMOTE Oversampling technique and conducted a comparison with and without





# A Comparative Study of Unsupervised Anomaly Detection Algorithms used in a Small and Medium-Sized Enterprise

Irina Petrariu<sup>1</sup>, Adrian Moscaliuc<sup>2</sup>, Cristina Elena Turcu<sup>3</sup>, Ovidiu Gherman<sup>4</sup>  
ASSIST Software SRL, Suceava, Romania<sup>1,2</sup>  
Stefan cel Mare University, Suceava, Romania<sup>3,4</sup>

**Abstract**—Anomaly detection finds application in several industries and domains. The anomaly detection market is growing driven by the increasing development and dynamic adoption of emerging technologies. Depending on the type of supervision, there are three main types of anomaly detection techniques: unsupervised, semi-supervised, and supervised. Given the wide variety of available anomaly detection algorithms, how can one choose which approach is most appropriate for a particular application? The purpose of this evaluation is to compare the performance of five unsupervised anomaly detection algorithms applied to a specific dataset from a small and medium-sized software enterprise, presented in this paper. To reduce the cost and complexity of a system developed to solve the problem of anomaly detection, a solution is to use machine learning (ML) algorithms that are available in one of the open-source libraries, such as the scikit-learn library or the PyOD library. These algorithms can be easily and quickly integrated into a low-cost software application developed to meet the needs of a small and medium-sized enterprise (SME). In our experiments, we considered some unsupervised algorithms available in PyOD library. The obtained results are presented, alongside with the limitations of the research.

**Keywords**—Unsupervised anomaly detection algorithms; small and medium-sized enterprise; traceability; open-source libraries

## I. INTRODUCTION

The current societal landscape has seen an increase in the quantity and complexity of information processed daily. Such increasing use is required for effective management of current industrial processes and depends on the data acquired from the process itself, data that is cleaned and converted into information that can be used to create meaningful visualizations, be fed in complex control and prediction algorithms, or even stored for future reference and use. Moreover, data reliability is paramount. Correct information must be used to obtain correct responses from the managed processes and incorrect information can lead to inefficiency, loss of precision, data, or product that in turn can negatively impact the organization's reputation or the bottom line.

In general, the data is acquired from the process via sensors, manually or through automated systems. To eliminate acquisition errors, data is sanitized and, if possible, corrected. This prevents the propagation of errors further in the system. Data that does not meet the criteria for correction may appear anomalous compared with its dataset values or regarding the

median of the dataset. In any scenario, this might indicate either erroneous data or valid data signaling a potential problem with data acquisition or in the process itself. Therefore, isolating anomalous data is an important indicator of data health and a promising path in data analysis.

Anomalies are unexpected instances of deviation from a large part of the dataset. Thus, solving them will allow for improving the efficiency of the underlying process [1]. In fact, according to various studies (e.g., [2]), applications based on anomaly detection could help an enterprise detect possible issues in time, before they emerge by identifying anomalous behavior, thus minimizing the risk of data loss and streamlining business processes. Anomaly detection finds application in multiple industries and domains, including healthcare, finance, manufacturing, construction, logistics, cyber security, and many others [3], [4]. There are various specific applications of anomalies detection, such as, system health monitoring, early detection of sepsis [5], event detection, product quality, intrusion detection, energy optimization, various real-time applications, to name only a few.

The anomaly detection market is witnessing growth, thus, according to [2], “the anomaly detection market size is expected to grow from USD 2.08 Billion in 2017 to USD 4.45 Billion by 2022, at a Compound Annual Growth Rate (CAGR) of 16.4%”. This growth is being driven by the increasing development and dynamic adoption of emerging technologies such as big data analytics, data mining and business intelligence, machine learning and artificial intelligence.

According to scientific literature, there are three main types of anomaly detection techniques, depending on the type of supervision: unsupervised, semi-supervised, and supervised. Essentially, the choice of anomaly detection method can be made according to the labels available in the dataset [6].

Considering the extensive variety of available anomaly detection algorithms, how can one choose which approach is most suitable for a particular application? Clearly, performance in anomaly detection is a significant factor in algorithm selection. Unfortunately, there is no one approach that is best in every context and for all domains. Depending on the specifics, one algorithm may be superior to the others for a given user or dataset. Selecting an appropriate algorithm for a specific application is still a difficult design choice [7]. This is even more important in traceability systems that must ensure that the

product's lifecycle is correct, where the detection of an anomaly in the process can have a major impact on the production pipeline. However, these workflows can vary from company to company, which means that the use of supervised learning would imply higher costs to develop a custom model, whereas unsupervised learning and more precisely, anomaly detection would allow building a model that does not require human intervention, does not need prior knowledge about the process and, at the same time, can be very dynamic in terms of feature selection.

Reviewing the literature on machine learning-based anomaly detection algorithms reveals the diversity of algorithms evaluation and comparison approaches used by researchers. Consequently, the authors of [7] draw attention to the inconsistency in splitting between training and test datasets, in the selection of performance metrics and in the threshold used to indicate anomalies. Moreover, they point out the ambiguity in the definition of the positive class (i.e., the class of interest) utilized to evaluate the various models. Because of these inconsistencies, the authors find it difficult to understand the experimental evaluations presented in different papers [7].

To reduce the cost and complexity of a system developed to solve the problem of anomaly detection, a solution is to use machine learning (ML) algorithms that are available in one of the open-source libraries, such as the scikit-learn library [8], [9] or the PyOD library [10]. These algorithms can be easily and quickly integrated into a low-cost software application developed to meet the needs of a small and medium-sized enterprise (SME). Once the relevant features considered for anomaly detection are selected and pre-processed, the integrated algorithms can be applied within the specific software application.

In this paper, we will examine the viability of implementing anomaly detection in a traceability system by applying unsupervised anomaly detection algorithms. In this regard, we will study and compare the performance of some of the unsupervised anomaly detection algorithms applied on a dataset provided by the information technology (IT) department of a small and medium-sized software enterprise. We considered some un-supervised algorithms available in one of the popular open-source libraries, namely PyOD library. This library provides several benefits over comparable existing libraries. For instance, it contains more than 20 algorithms, it “implements combination methods for merging the results of multiple detectors and outlier ensembles which are an emerging set of models”, and “all models are covered by unit testing with cross platform continuous integration, code coverage and code maintainability checks” [10]. These benefits have led to its widespread adoption in academic and commercial applications [10]. According to [11], [12], the GitHub repository has more than 10,000 monthly visitors, and more than 6,000 monthly downloads for PyPOD.

The remainder of this paper is structured as follows. Section II provides an overview of machine learning-based solutions for traceability domain, methods and algorithms in anomaly detection and the evaluated anomaly detection algorithms. The methodology we relied on to conduct the presented research is also discussed. Section III describes the

experimental process and results obtained on a real dataset. Section IV is dedicated to presenting insights on the performance of the algorithms. The limitations and directions for future research are presented in Section V. The final section provides the concluding remarks of the paper.

## II. MATERIALS AND METHODS

In this section, we review prior work in terms of machine learning in traceability, and methods and algorithms in anomaly detection.

### A. Machine Learning in Traceability

By employing traceability systems, products can have better quality, or the workflow can be improved. Thus, this concept has been applied in a variety of domains, ranging from managing the food supply chain [13], [14] to the automotive industry [15]. Moreover, given the high volume of data, ML algorithms could be used to analyze and provide relevant information that can be used in the decision-making process.

Given that food safety is a critical concern, traceability has been used in this industry in order to follow the process taken by perishables to ensure safety and quality. For example, De Nadai Fernandes et al. [16] employed three supervised ML algorithms to determine the source of bovine meat in Brazil, where there was a loss of information during the slaughtering or marketing processes. On the other hand, Alfian et al. [13] used Radio Frequency IDentification (RFID) tags and Internet of Things (IoT) sensors to collect information regarding the environment and track when produce would pass through a space, for example, a warehouse door. They also employed supervised learning to identify the direction of the product and, thus, determine whether the products were safely stored. To prevent food safety incidents, Wang et al. [14] developed a traceability system that ensured quality at each stage of the production pipeline by developing supervised ML algorithms to determine the quality at each stage and establish the final quality, whereas Shahbazi and Byun [17] utilized ML and blockchain to detect counterfeits and ensure the validity of the expiration dates.

Sharma et al. [18] reviewed the use of ML algorithms in the agricultural supply chain and observed that these algorithms were implemented at each step of the production process to improve efficiency. For instance, in the preproduction stage, ML algorithms were used to predict the harvest, soil properties and irrigation management. In the production step, they were used to predict the weather, protect the harvest, detect weeds, manage animals, and overview the harvest quality. The processing step consisted of algorithms used to predict the demand and plan the production, while in the distribution phase they were used to improve transportation, analyze the consumers, and manage the inventory.

Other domains have also included ML algorithms to analyze performance, for instance, in the automotive industry [15], or to determine validity in software maintenance for traceability link recovery [19]. One important observation is the focus on quality control and preventing issues that could occur. However, most of the discussed systems used supervised machine learning algorithms that need labeled data as input, which means that the process must be firmly

established, and any change means having to reformat and revalidate the training dataset. In this context, we looked at unsupervised anomaly detection algorithms that could be applied to a more dynamic process to alert the user of any abnormalities in the system by using an unlabeled dataset.

The next sections provide brief descriptions of the algorithms used for anomaly detection.

### B. Methods and Algorithms in Anomaly Detection

Detection of anomalous instances in various datasets is of great importance in many processes. Outlier observations, records, recorded values, states, and devices can either affect the workflow of processes or can induce bias in computed values or scores. Removing the outliers is a valid and known technique, but the detection of the aforementioned anomalous values is not an easy process – especially when the data is not identified and tagged [6].

Anomaly detection techniques are employed in various working domains, but especially in supply chain management, where the capabilities of various techniques allow to manage complicated processes, using predictive algorithms and other use cases. For example, using blockchain technologies in supply chain systems allows for novel methods to manage all aspects of the process. However, to have a robust data model and verification mechanism to ensure the integrity of the processes it is required to implement mechanisms to correct anomaly signals in the acquired data required in verification processes for the business logic involving transactions (both resource intensive and critical for the wellbeing of the platform) [20], [21] between entities in the platform, for a better quality in supply chain management (not only from a performance point of view, but also from a security perspective).

In many applications that require anomaly data detection, ML algorithms are used to highlight the relevant data for future correction, removal, or analysis [22]. Many times, the detection algorithms are paired with traditional detection systems, usually rule-based, for better performance. In this regard, there are various application domains where these types of algorithms are used [6]: network security for intrusion detection (used for behavior analysis in enterprise settings for known and novel threats), surveillance for suspicious moves and actions (via visual and audio capture systems) [23], [24], detection of fraudulent transactions in banking industry (including transactions involving digital goods) [25], [26], energy optimization in smart buildings [27], medical smart equipment (capable of identification and analysis of anomalies to assist in medical diagnosis) [28], [29], and, generally, in use-cases where anomalous states can appear infrequently enough in the operating processes to be properly treated, but pose enough dangers to warrant such a system.

Even in processes that are tied to physical mediums, like, for example, nuclear radiation detection [30], telemetry data for spacecraft operations [31], traffic patterns analysis [32], sensor arrays and IoT systems [33], unmanned ground and aerial vehicles detection [34], edge computing systems and novel large-scale IT systems [35] or network quality of service assurance by using a Greedy algorithm [36] or even genetic

algorithms [37], such ML algorithms can be used for identification of anomalies. Moreover, similar algorithms are used in domains like supply chain management, where genetic rule-based and graph-based detection methods are employed to verify business transactions regarding their validity [20].

Anomaly detection domain can be classified in three types [6], based on the approach regarding labelling of the dataset content: supervised anomaly detection where the training data and the test data are fully labelled (in practice this is less used given that labelling the data is not always feasible or even possible), semi-supervised anomaly detection where the training is done on non-anomalous datasets (the anomalies being detected when they deviate from the “correct” model) and unsupervised anomaly detection that does not require labels to classify data (most flexible approach), the distinction being made on the internal properties of the dataset.

As stated in the introduction section, this paper has considered the evaluation of several unsupervised algorithms for anomaly detection in the context of traceability. According to [6], unsupervised anomaly detection algorithms can be roughly classified into the following main categories as illustrated in Fig.1 (1) Nearest-neighbor based techniques, (2) Clustering-based methods, (3) Statistical algorithms, (4) Subspace techniques, (5) Classifier-based algorithms.

In this paper, the evaluated algorithms are part of the first three groups.

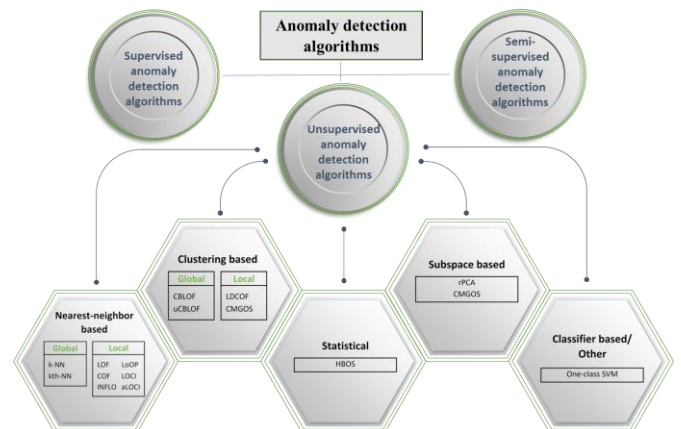


Fig. 1. Taxonomy of Unsupervised Anomaly Detection Algorithms (Adapted after [6]).

### C. The Evaluated Anomaly Detection Algorithms

In the following, it briefly presents the five anomaly detection algorithms that were evaluated. The k-nearest neighbor (k-NN or KNN) algorithm consists of finding the nearest neighbors and calculating the anomaly score based either on the distance to the nearest neighbor [38] or the mean distance to the k nearest neighbors [39]. One of the drawbacks of this algorithm is that it detects only global anomalies, an issue that was tackled in the Local Outlier Factor (LOF) algorithm [40], which was the first algorithm to detect local anomalies. The Cluster-Based Local Outlier Factors (CBLOF) algorithm [41] uses the grouping to determine the dense areas from the data and uses a heuristic to classify the groups,

whereas Histogram-based Outlier Score (HBOS) [42] is based on statistics and assumes that there are no dependencies between the features of the model. One challenge faced by clustering algorithms is choosing the number of groups, which was addressed by Local Correlation Integral (LOCI) [43] that uses a maximization approach.

The implementation of these algorithms in various software libraries, such as the PyOD library [10], facilitates their use in the development of software applications for anomaly detection.

#### D. Performance Criteria

In this paper, we have examined the possibility of employing anomaly detection in a traceability system by applying five unsupervised anomaly detection algorithms and compared their performance.

The algorithms evaluated in the considered scenario must be analyzed from a performance perspective. Given that not necessarily all algorithms can be implemented in the traceability platform (for performance reasons), usually the best algorithm will be employed in the final product, for best performance/accuracy ratio. Alternatively, two or three algorithms can be employed in certain circumstances, where their performance can compensate for their weaknesses in certain datasets configurations or certain limited cases. Thus, it is important to establish the performance of the potential solution in the given circumstances, including when adjusting the settings in the classifier model.

In this regard, a criterion that is often used to test the performance of algorithms in machine learning is the ROC curve (Receiver Operating Characteristics) [44]. ROC metric will help establish the performance of the model (higher the value, better the outcome) by plotting the rate of true positives compared (higher is better) with the rate of false positives (lower is better) and thus establishing a threshold for the performance of the model in classifying the input data [44]; this approach is important when deciding between various algorithms or when adjusting operating parameters of a given algorithm.

Another important criterion is the accuracy score. In performance metrics [45], the accuracy score is the measure by which the classifier will offer correct predictions compared with the total number of predictions made. Obviously, a greater accuracy is a highly desirable behavior of the model.

Finally, precision @ rank n represents the precision of the model up to n<sup>th</sup> prediction from the total number of predictions [46]. This metric helps usually in choosing a better model for a given type of problem, given that the goal is to obtain a good fit for our algorithm.

$$\text{Precision}@k = \frac{\text{true\_positives}@k}{\text{true\_positives}@k + \text{false\_positives}@k} \quad (1)$$

Training time is also an important factor in choosing a certain algorithm. Given that the time spent training the models can be an expensive proposition (for example in big datasets or when the datasets change in time, requiring re-training of the model), choosing an algorithm that is faster on training is better, as long as the performance metrics are not suffering.

As highlighted in [7], different splits of the training and test data can be used while comparing the performances of various algorithms. Thus, decisions have to be made regarding the splitting of training and test sets. For anomaly detection, it must additionally be decided whether one of these two sets will contain normal data, abnormal data, or both. Regardless of the decision taken, it must be used in a consistent manner when evaluating different algorithms [7].

In the next section, we present the considered use case.

#### E. Use Case

In order to perform the proposed comparison of some anomaly detection algorithms, we have considered the use case of the traceability of different equipment types in an IT department Fig. 2 presents the main steps of our process of evaluation.

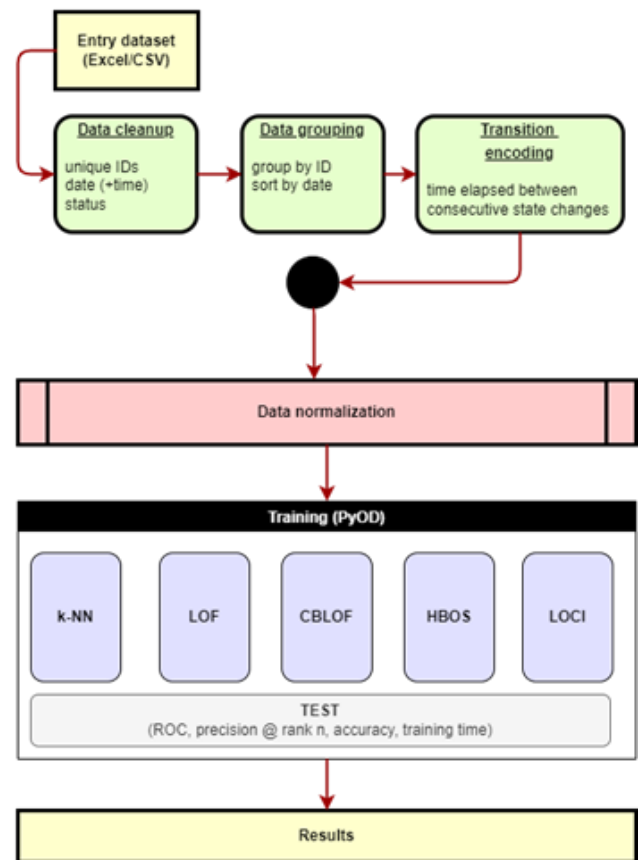


Fig. 2. The Workflow of the Experiment.

Next, we describe the data pre-processing and applied methods.

To perform the proposed anomaly detection comparison, we were provided with a set of data in the form of an Excel spreadsheet or a CSV (comma separated values) file that contained information regarding the management of equipment in an IT department. The spreadsheet was composed of records of items that were registered and then assigned to employees. There were cases when items malfunctioned, so they were sent back to the IT department for repairs and assigned back to an employee or had to be scrapped. The file consists of multiple

data columns, such as unique ID, date, equipment name, equipment type, employee name, status, etc. The dataset contained a number of 347 transitions for 130 items with 130 records for the acquired status, 177 instances of assigning an item to an employee for use, 22 cases of sending the item to the IT department for repairs, and 18 recorded for items being scrapped. It was considered that the data provided was valid and could be used for training a machine learning model, and that artificial anomalies could be added to the set of data for validation. It should be noted that this process can be applied to other types of items, regardless of their statuses as this process is generic and does not require any previous setup.

The first step of the pre-processing phase is extracting the relevant information, in this case from the Excel or CSV file, but other sources, such as databases, can be used. The significant attributes are the item unique identifier, and the date and status of the item when that record was made. Depending on how the data is stored, the status of the item may have a variety of formats, ranging from integers to strings. For example, we considered the coding of statuses presented in Table I.

Based on the previous coding of the statuses, Table II, illustrates the case of a monitor with the ID 132606, which was purchased (status with unique ID 1) on January 27<sup>th</sup> and given in use (status with unique ID 3) to an employee the next day, on January 28<sup>th</sup>. On June 25 of the same year, the item identified with ID 132606 is sent for repairs.

The date, and time if available, must have the same format and should be converted to a format that would allow for simple time difference calculations. Since the provided spreadsheet contained only the registration date, it was converted to the number of days since January 1st, 1900, in accordance with the ISO 8601:2000 YYMMDD format.

Secondly, to define the item transitions, all data must be grouped according to the unique identifier and sorted by date inside that group. If time is known, it should also be considered in the ordering, and this would be of greater significance in situations in which items transition multiple statuses on the same day. In our case, after going through these steps, the data for an item should reveal how an item was purchased and then assigned to employees with some cases where it was sent for repairs or was scrapped. A transition is represented based on the time elapsed between the current state and the previous state in the form of a number of days, or a number of minutes or milliseconds (depending on the granularity of the data), if the data contained time. The transition token is composed of the concatenation of the current and previous status identifier, which was defined in the previous step. A sample of the data is shown in Table III.

In this example, if item 132606 was purchased on the 27<sup>th</sup> of January and assigned (status with the three unique ID) to an

employee on the 28<sup>th</sup> of January that same year, then this transition will be characterized by a one-day time interval, resulting from the difference between the two dates, and the “13” token, which represents the concatenation of the unique identifiers assigned to the status attributes. These will be the two features that will be used to train the ML model, which were selected as relevant as a result of an analysis of what information is critical in a traceability system with the purpose of creating a trained model that could point out the incorrect transitions. Therefore, mistakes could be made while changing the status of an item, for instance, moving a monitor from acquired to scrapped would be invalid, whereas if, for example, the monitor is left in the in-repair state for a prolonged period reveals a different type of issue such as the lack of available employees to check the monitor. When an item is added to the system and is assigned its first status, the elapsed time should be set to zero and the token should consist of the doubling of the unique identifier status, in our case, the token for the first transition of item 132606 is “11”. If an item starts with a different status, then the ML algorithm should be able to signal it as an anomaly.

TABLE I. THE CODING OF THE CONSIDERED STATUSES

Status	Code
Purchased	1
Scrapped	2
In use	3
Maintenance	4

TABLE II. SAMPLE OF THE ENCODED DATA RECEIVED FROM IT DEPT.

Unique item ID	Status	Date
132606	1	January 26, 2021
132606	3	January 28, 2021
132606	4	June 25, 2021
132606	3	July 02, 2021
134338	1	April 29, 2021
134338	3	May 14, 2021
134338	2	July 30, 2021

TABLE III. EXAMPLE OF ITEM TRANSITIONS

Unique item ID	Transition token	Timelapse (days)
132606	11	0
132606	13	1
132606	34	148
132606	43	7
134338	11	0
134338	13	15
134338	32	77



The third step before training the models is normalizing the data, which in this case was performed using the preprocessing scale function from the scikit-learn [9] Python package. Normalization needs to be applied to each feature, because having overly broad scales could cause issues when training the ML model. This is avoided through the fact that the new values are smaller and at the same time they maintain the general distribution and data ratios, which means that the relevant information is preserved. Before applying this step, we also added a number of four artificial anomalies consisting of invalid transitions, new statuses, and prolonged periods of time, which should be signaled by the ML model. The added artificial anomalies had to be run through the same scaling process as the valid set of data to avoid inaccuracies. We added such a small number of anomalies, because, as per their definition, anomalies are events that occur rarely.

To analyze the performance of the applied algorithms the data has been formatted and split into a file that contains all the valid transitions, a file that contains 80% of the data and four artificial anomalies that can be used for training the model, and a file that contains the rest of the 20% of the data and other four artificial anomalies that can be used as the test data, which will be utilized to validate whether the model is overfitted. The split between the test and training data is done randomly by shuffling the item IDs and then selecting 20% of the IDs and their corresponding transitions for the test data. The rest of the items are assigned to the training dataset. In addition, when the data is split, it should be normalized separately to avoid any data leakage. Given that the IDs are separated randomly, this means that the data will be split differently when the process is run again on the same set of data.

### III. RESULTS

Next, the results of the training and tests processes for the ML algorithms having as input the datasets described in the previous section of this paper are presented. To evaluate and compare the anomaly detection methods we use the standard metrics of precision, ROC and accuracy. Also, it was considered the training time for each algorithm. The experimental evaluation is conducted on a laptop with i7-8550U CPU @1.80 GHz x 8, 16 GB of RAM, running Ubuntu 16.04 LTS. The code is written in Python 3.

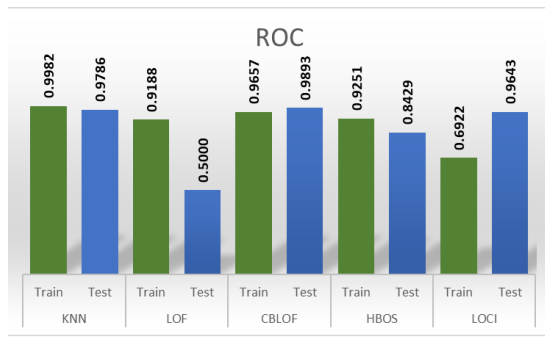
In the experiments conducted, the split of data was performed considering normal and anomalous data, both in the training and in the testing phase. For the first analysis, we used the training dataset that had 80% of the whole data, and the test train dataset composed of 20% of the provided information. Each set contained the same four artificial anomalies that were used for validation. The first set of data was used to train five separate anomaly detection algorithms available in the PyOD Python package [10] by first initializing them and then applying the fit function that received the training set of data. This resulted in a list of prediction labels and outlier scores of the training data, which we used to compute the accuracy of the model based on the training dataset. To evaluate the model, the predict and *decision\_function* functions were used, which received the test dataset as input, whose results were used to calculate the accuracy of the test predictions. Accuracy was calculated based on the assumption that the received dataset contained only valid records and that there were four artificial anomalies. To calculate the ROC and the Precision @ rank n, we used the *evaluate\_print* function provided by the PyOD library. It was also recorded the time needed to train each model. Table IV contains the results of these metrics for the analyzed five algorithms. The evaluation results are shown in Fig. 3.

Given that the provided dataset was small, during the second analysis the complete dataset was used as the training data and some predictions were made based on four cases: normal record for item just being added to the system (N1), normal record for item being assigned to employee (N2) that was taken from the initial dataset, anomaly with non-existent states (A1), and anomaly with item assigned to employee but with an anomalous time (A2). Before sending these datasets to the algorithms for predictions they were processed by using the same normalization parameters used for the training set. Table V shows the results of training the five algorithms with the full set of data and using the previously described cases to validate the predictions. The evaluation results are shown in Fig. 4.

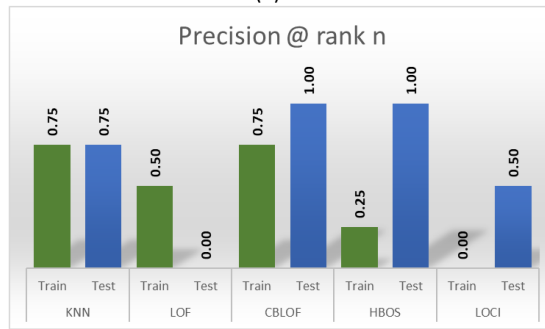
Nevertheless, in a normal setting, the prediction would be made based on a model that was trained using the full set of data, which could occur right when there is a request to predict whether a transition is anomalous or not, if the training is fast enough, or periodically, in which case it would contain only the records that were registered until the time of training.

TABLE IV. RESULTS OF ALGORITHM BASED ON TRAINING AND TEST DATASET

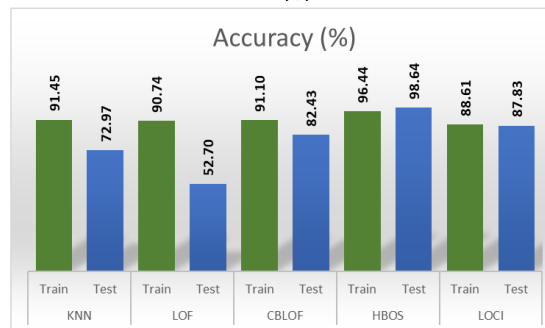
Metric/ algorithm	KNN		LOF		CBLOF		HBOS		LOCI	
	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
ROC	0.9982	0.9786	0.9188	0.5000	0.9657	0.9893	0.9251	0.8429	0.6922	0.9643
Precision @ rank n	0.7500	0.7500	0.5000	0.0000	0.7500	1.0000	0.2500	1.0000	0.0000	0.5000
Accuracy (%)	91.45	72.97	90.74	52.70	91.10	82.43	96.44	98.64	88.61	87.83
Time (seconds)	0.0023		0.0037		0.0699		0.0019		44.8992	



(a)



(b)



(c)

Fig. 3. Experiment Results for the First Analysis: (a) ROC Curve; (b) Precision @ Rank n; (c) Accuracy (Expressed as Percentages).

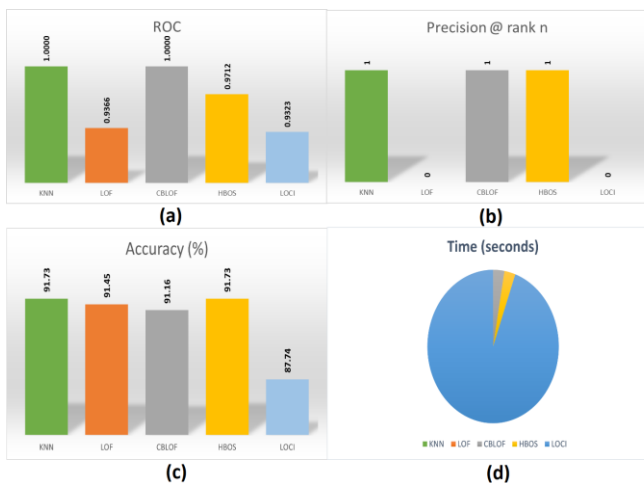


Fig. 4. Experiment Results for the Second Analysis: (a) ROC Curve; (b) Precision @ Rank n; (c) Accuracy (Expressed as Percentages); (d) Execution Time (in Seconds).

TABLE V. RESULTS OF THE ALGORITHMS TRAINED ON THE FULL DATASET AND VALIDATED BASED ON FOUR PREDICTIONS

Metric/algorithm	KNN	LOF	CBLOF	HBOS	LOCI
ROC	1.0000	0.9366	1.0000	0.9712	0.9323
Precision @ rank n	1.0000	0.0000	1.0000	1.0000	0.0000
Accuracy (%)	91.73	91.45	91.16	91.73	87.74
Time (seconds)	0.0042	0.0044	2.5900	2.5399	88.51
N1 prediction	Correct	Correct	Correct	Correct	Correct
N2 prediction	Correct	Correct	Correct	Wrong	Correct
A1 prediction	Correct	Correct	Correct	Correct	Wrong
A2 prediction	Correct	Correct	Correct	Correct	Wrong

#### IV. DISCUSSION

In this section we discuss and analyze the results of running the five machine learning algorithms, mainly k-NN (k-nearest-neighbor) - a nearest-neighbor-based unsupervised algorithm focused on detection of global anomalies (global relative to the dataset) with low computational impact [47], LOF (Local Outlier Factor) - a nearest-neighbor-based algorithm able to detect local anomalies alongside global ones [40], [48], LOCI (Local Correlation Integral) - a nearest-neighbor-based local algorithm with increased precision over k-NN but also with increased computational complexity [43], [45], CBLOF (cluster-based local outlier factor) - a clustering-based global algorithm [49] and HBOS (histogram-based outlier score), a very fast statistical algorithm almost an order of magnitude faster than k-NN [50].

By examining the results for the ROC scores from Table IV, it can be observed that most of the algorithms had good outcomes for the training values apart from LOCI with 0.6922, with the best being KNN (0.9982) followed by CBLOF (0.9657). On the other hand, the ROC values for the test sets show us that not all the models generalized well, for instance, the test ROC for LOF was 0.5, whereas HBOS had a lower score than the other algorithms with 0.8429. The best results were achieved again by KNN (0.9786) and CBLOF (0.9893), followed by LOCI (0.9643).

In terms of the precision @ rank n score, overall, CBLOF had the best results (0.75 for training and 1.0 for test) followed by KNN (0.75 for both the training and test datasets). Although HBOS had a low value of 0.25 for the training set, it had an exceptionally good score of 1.0 for the test set, whereas the other algorithms had low scores in general with 0.5 and 0.0 for train and test, respectively, for LOF, and 0.0 and 0.5 for train and test, respectively, for LOCI.

The accuracy score results revealed that HBOS performed the best with 96.44% for training and 98.64% for test with CBLOF being second with 91.10% for training and 82.43% for test. Even though KNN had a good result for the training dataset, 91.45%, it scored lower for the test dataset, 72.97%. Albeit having lower accuracy ratings, LOCI had similar values for both the train and test sets with 88.61% and 87.83%, respectively, whereas LOF had a considerable difference

between the test and train scores (90.74% and 52.70% respectively), which denotes that the model did not generalize.

Although the training time would not be a major factor to consider if the training is performed daily when there is no heavy traffic in the system, it can be still noted that LOCI had a significantly higher training time compared to the other algorithms. Its training lasted almost 45 seconds even if the dataset was relatively small while all the other models had times lower than a second, making LOCI definitively not a suitable candidate for such a system.

The second round of experiments displayed in Table V, which are closer to a real scenario, revealed some interesting results. Firstly, when using the full dataset for training, all algorithms had high scores for the ROC values with KNN and CBLOF having 1.0 followed by LOF and LOCI that had remarkably similar results, 0.9366 and 0.9323, respectively. On the other hand, the precision @ rank n was either exceptionally good with 1.0 for KNN, HBOS and CBLOF or bad with 0.0 for LOF and LOCI.

The accuracy scores were also higher in general, with KNN and HBOS having the same outcome of 91.73% followed by LOF at 91.45%, CBLOF at 91.16% and LOCI at 87.74%, which had a very similar result to the first experiment denoting a consistent pattern in the ability of this model to predict the anomalies for this dataset. In terms of the training time, KNN and LOF had the lowest results with times under a second. However, CBLOF and HBOS had significantly higher times (around 2.5 seconds), which can indicate a more rapid increase in time given that the training from the first experiment was performed on 80% of the data. LOCI still had the highest time with 88.51 seconds.

Regarding the four predictions, two normal cases and two anomalies, KNN, LOF and CBLOF correctly predicted all four cases, HBOS wrongly detected N2 as an anomaly, whereas LOCI was not able to detect the two anomalies. Although HBOS had similar results to KNN and CBLOF for the other evaluation conditions, it did not perform as well in terms of the test prediction. Thus, overall, the best results were achieved by KNN followed closely by CBLOF apart for the training time.

## V. LIMITATIONS

The size of the used dataset is a limiting factor in our work, however, even with such a small size we were able to demonstrate that anomaly detection can be applied to traceability with good results. Nonetheless, having a larger dataset could offer more insight, an analysis that could be undertaken in future work, where we could also include more algorithms in the comparison. The difficulty in creating a database for anomaly detection lies in the fact that the results will emulate the logic that was used in generating the data, thus, it is important to have access to a real dataset.

In future research, data splitting will be performed considering the normal data for training and all anomalous samples in the test set.

In terms of the discussed logic, adding a new process or status will automatically result in an anomaly detection. To solve this problem, the described logic could be adapted to add

a threshold, for example, there must be a minimum number of products that went through a status in order to send that transition to the anomaly detection algorithm. Another option would be to check if a generated group has lower elements than a set threshold. However, this could also mean that transitions that could be anomalous are not sent or are not taken into account by model. Handling this issue could be further investigated in future work.

In this analysis, only one process was considered. In order to improve the performance of the model, anomaly detection should be conducted for each type of process if the company workflow has various processes with different statuses. Obviously, for training a model there must be a minimum number of transitions for each category, which might be determined by user input or by involving a user in model validation. Once the model was able to properly detect anomalies, the threshold could be automatically set for future categories. Nevertheless, this could open the door to semi-supervised learning, which will be investigated in future work.

## VI. CONCLUSIONS

In recent years, there has been an increase in the number of artificial intelligence-based solutions for problems in various fields. Anomaly detection is an issue for which there are currently several approaches. One of the most widespread methods involves the use of ML techniques.

In this paper, the performance of five unsupervised anomaly detection algorithms regarding a traceability dataset that contains information on the management of devices from an IT department was analyzed and compared. The models used with anomaly detection algorithms based on machine learning do not require labeled data. Given the importance of reproducibility in research, we presented all the information regarding the implementation that allow double-checking the results and verifying whether they are reliable. By analyzing the precision, accuracy, ROC, and time we determine which algorithms tend to perform better or worse on the presented use case. We have demonstrated experimentally that these algorithms can be successfully applied to determine whether a new transition is an anomaly with an accuracy of up to 91.73%.

## ACKNOWLEDGMENT

This research was funded by the project “119722/Centru pentru transferul de cunoștințe către întreprinderi din domeniul ICT—CENTRIC, Contract subsidiar 15568/01.09.2020, Smart Tracking Platform (STP)”, contract no. 5/AXA 1/1.2.3/G/13.06.2018, cod SMIS 2014+ 119722 (ID P\_40\_305).

## REFERENCES

- [1] Y. Wang et al., “Iterative anomaly detection,” in 2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Jul. 2017, pp. 586–589. doi: 10.1109/IGARSS.2017.8127021.
- [2] “Anomaly detection market by Solution (Network and user behavior anomaly detection), technology (Big data analytics, data mining and business intelligence, machine learning and artificial intelligence), deployment, service, vertical - Global forecast to 2022,” MarketsandMarkets, Market report. [Online]. Available: <https://www.marketsandmarkets.com/Market-Reports/anomaly-detection-market-138133262.html>.

- [3] M. Bahri, F. Salutari, A. Putina, and M. Sozio, "AutoML: State of the art with a focus on anomaly detection, challenges, and research directions," *Int. J. Data Sci. Anal.*, vol. 14, no. 2, pp. 113–126, Aug. 2022, doi: 10.1007/s41060-022-00309-0.
- [4] I. K. Nti, A. F. Adekoya, B. A. Weyori, and O. Nyarko-Boateng, "Applications of artificial intelligence in engineering and manufacturing: a systematic review," *J. Intell. Manuf.*, vol. 33, no. 6, pp. 1581–1601, Aug. 2022, doi: 10.1007/s10845-021-01771-6.
- [5] L. Begic Fazlic et al., "A novel hybrid methodology for anomaly detection in time series," *Int. J. Comput. Intell. Syst.*, vol. 15, no. 1, p. 50, Jul. 2022, doi: 10.1007/s44196-022-00100-w.
- [6] M. Goldstein and S. Uchida, "A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data," *PLoS One*, vol. 11, no. 4, 2016, doi: <https://doi.org/10.1371/journal.pone.0152173>.
- [7] M. Alvarez, J.-C. Verdier, D. K. Nkashama, M. Frappier, P.-M. Tardif, and F. Kabanza, "A revealing large-scale evaluation of unsupervised anomaly detection algorithms," *ArXiv Prepr. ArXiv220409825*, 2022.
- [8] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, Jan. 2012.
- [9] D. Cournapeau, "scikit-learn," *scikit-learn*, 2022. <https://scikit-learn.org/stable/about.html> (accessed Jul. 24, 2022).
- [10] Y. Zhao, Z. Nasrullah, and Z. Li, "PyOD: A Python toolbox for scalable outlier detection," *J. Mach. Learn. Res.*, vol. 20, no. 96, pp. 1–7, 2019, [Online]. Available: <http://jmlr.org/papers/v20/19-011.html>.
- [11] Y. Zhao and M. K. Hryniewicki, "DCSO: Dynamic combination of detector scores for outlier ensembles," *ArXiv Prepr. ArXiv191110418*, 2019, doi: <https://doi.org/10.48550/arXiv.1911.10418>.
- [12] Y. Zhao, Z. Nasrullah, M. K. Hryniewicki, and Z. Li, "LSCP: Locally selective combination in parallel outlier ensembles," in *Proceedings of the 2019 SIAM International Conference on Data Mining*, 2019, pp. 585–593.
- [13] G. Alfian et al., "Improving efficiency of RFID-based traceability system for perishable food by utilizing IoT sensors and machine learning model," *Food Control*, vol. 110, p. 107016, Apr. 2020, doi: 10.1016/j.foodcont.2019.107016.
- [14] J. Wang, H. Yue, and Z. Zhou, "An improved traceability system for food quality assurance and evaluation based on fuzzy classification and neural network," *Food Control*, vol. 79, pp. 363–370, Sep. 2017, doi: 10.1016/j.foodcont.2017.04.013.
- [15] M. Syafrudin, G. Alfian, N. L. Fitriyani, and J. Rhee, "Performance analysis of IoT-based sensor, big data processing, and machine learning model for real-time monitoring system in automotive manufacturing," *Sensors*, vol. 18, no. 9, Art. no. 9, Sep. 2018, doi: 10.3390/s18092946.
- [16] E. A. De Nadai Fernandes, G. A. Sarriés, M. A. Bacchi, Y. T. Mazola, C. L. Gonzaga, and S. R. V. Sarriés, "Trace elements and machine learning for Brazilian beef traceability," *Food Chem.*, vol. 333, p. 127462, Dec. 2020, doi: 10.1016/j.foodchem.2020.127462.
- [17] Z. Shahbazi and Y.-C. Byun, "A procedure for tracing supply chains for perishable food based on blockchain, machine learning and fuzzy logic," *Electronics*, vol. 10, no. 1, Art. no. 1, Jan. 2021, doi: 10.3390/electronics10010041.
- [18] R. Sharma, S. S. Kamble, A. Gunasekaran, V. Kumar, and A. Kumar, "A systematic literature review on machine learning applications for sustainable agriculture supply chain performance," *Comput. Oper. Res.*, vol. 119, p. 104926, Jul. 2020, doi: 10.1016/j.cor.2020.104926.
- [19] C. Mills and S. Haiduc, "A machine learning approach for determining the validity of traceability links," in *2017 IEEE/ACM 39th International Conference on Software Engineering Companion (ICSE-C)*, May 2017, pp. 121–123, doi: 10.1109/ICSE-C.2017.86.
- [20] B. Oh, T. J. Jun, W. Yoon, Y. Lee, S. Kim, and D. Kim, "Enhancing trust of supply chain using blockchain platform with robust data model and verification mechanisms," in *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)*, Oct. 2019, pp. 3504–3511, doi: 10.1109/SMC.2019.8913871.
- [21] M. Khalfaoui, R. Molva, and L. Gomez, "Secure alert tracking in supply chain," in *2013 International Conference on Security and Cryptography (SECURITY)*, Jul. 2013, pp. 1–11.
- [22] S. B. Wankhede, "Anomaly detection using machine learning techniques," in *2019 IEEE 5th International Conference for Convergence in Technology (I2CT)*, Mar. 2019, pp. 1–3, doi: 10.1109/I2CT45611.2019.9033532.
- [23] L. Kratz and K. Nishino, "Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2009, pp. 1446–1453, doi: 10.1109/CVPR.2009.5206771.
- [24] P. Khaire and P. Kumar, "A semi-supervised deep learning based video anomaly detection framework using RGB-D for surveillance of real-world critical environments," *Forensic Sci. Int. Digit. Investig.*, vol. 40, p. 301346, Mar. 2022, doi: 10.1016/j.fsidi.2022.301346.
- [25] V. Chang, L. M. T. Doan, A. Di Stefano, Z. Sun, and G. Fortino, "Digital payment fraud detection methods in digital ages and Industry 4.0," *Comput. Electr. Eng.*, vol. 100, p. 107734, May 2022, doi: 10.1016/j.compeleceng.2022.107734.
- [26] J. Vanhoeyveld, D. Martens, and B. Peeters, "Value-added tax fraud detection with scalable anomaly detection techniques," *Appl. Soft Comput.*, vol. 86, p. 105895, Jan. 2020, doi: 10.1016/j.asoc.2019.105895.
- [27] Y. Himeur, K. Ghanem, A. Alsalemi, F. Bensaali, and A. Amira, "Artificial intelligence based anomaly detection of energy consumption in buildings: A review, current trends and new perspectives," *Appl. Energy*, vol. 287, p. 116601, Apr. 2021, doi: 10.1016/j.apenergy.2021.116601.
- [28] N. Melnykova, R. Kulievych, Y. Vyclus, K. Melnykova, and V. Melnykov, "Anomalies detecting in medical metrics using machine learning tools," *Procedia Comput. Sci.*, vol. 198, pp. 718–723, Jan. 2022, doi: 10.1016/j.procs.2021.12.312.
- [29] J. Lin, E. Keogh, A. Fu, and H. Van Herle, "Approximations to magic: finding unusual medical time series," in *18th IEEE Symposium on Computer-Based Medical Systems (CBMS'05)*, Jun. 2005, pp. 329–334, doi: 10.1109/CBMS.2005.34.
- [30] P. Zhou and S. Abbaszadeh, "Towards real-time machine learning for anomaly detection," in *2020 IEEE Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC)*, Oct. 2020, pp. 1–3, doi: 10.1109/NSS/MIC42677.2020.9507937.
- [31] M. M. Fernández, Y. Yue, and R. Weber, "Telemetry anomaly detection system using machine learning to streamline mission operations," in *2017 6th International Conference on Space Mission Challenges for Information Technology (SMC-IT)*, Sep. 2017, pp. 70–75, doi: 10.1109/SMC-IT.2017.19.
- [32] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Comput. Surv.*, vol. 41, no. 3, p. 15:1-15:58, Jul. 2009, doi: 10.1145/1541880.1541882.
- [33] L. Erhan et al., "Smart anomaly detection in sensor systems: A multi-perspective review," *Inf. Fusion*, vol. 67, pp. 64–79, Mar. 2021, doi: 10.1016/j.inffus.2020.10.001.
- [34] S. Khan, C. F. Liew, T. Yairi, and R. McWilliam, "Unsupervised anomaly detection in unmanned aerial vehicles," *Appl. Soft Comput.*, vol. 83, p. 105650, Oct. 2019, doi: 10.1016/j.asoc.2019.105650.
- [35] O. M. Ezeme, Q. H. Mahmoud, and A. Azim, "A deep learning approach to distributed anomaly detection for edge computing," in *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, Dec. 2019, pp. 992–999, doi: 10.1109/ICMLA.2019.00169.
- [36] Ç. Ateş, S. Özdel, M. Yıldırım, and E. Anarım, "Network anomaly detection using header information with greedy algorithm," in *2019 27th Signal Processing and Communications Applications Conference (SIU)*, Apr. 2019, pp. 1–4, doi: 10.1109/SIU.2019.8806451.
- [37] Q. Su and J. Liu, "A network anomaly detection method based on genetic algorithm," in *2017 4th International Conference on Systems and Informatics (ICSAI)*, Nov. 2017, pp. 1029–1034, doi: 10.1109/ICSAI.2017.8248437.
- [38] M. Hassan, H. Maher and K. Gouda, "A Fast and Efficient Algorithm for Outlier Detection Over Data Streams," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 11, 2021.

- [39] F. Angiulli and C. Pizzuti, "Fast outlier detection in high dimensional spaces," in *Principles of Data Mining and Knowledge Discovery*, Berlin, Heidelberg, 2002, pp. 15–27. doi: 10.1007/3-540-45681-3\_2.
- [40] M. U. Rehman and D. M. Khan, "Local Neighborhood-based Outlier Detection of High Dimensional Data using different Proximity Functions," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 4, 2020.
- [41] Z. He, X. Xu, and S. Deng, "Discovering cluster-based local outliers," *Pattern Recognit. Lett.*, vol. 24, no. 9, pp. 1641–1650, Jun. 2003, doi: 10.1016/S0167-8655(03)00003-5.
- [42] M. Goldstein and A. Dengel, "Histogram-based Outlier Score (HBOS): A fast unsupervised anomaly detection algorithm.", KI-2012: Poster and Demo Track, 2012.
- [43] S. Papadimitriou, H. Kitagawa, P. B. Gibbons, and C. Faloutsos, "LOCI: Fast outlier detection using the local correlation integral," in *Proceedings 19th International Conference on Data Engineering (Cat. No.03CH37405)*, Mar. 2003, pp. 315–326. doi: 10.1109/ICDE.2003.1260802.
- [44] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, Jun. 2006, doi: 10.1016/j.patrec.2005.10.010.
- [45] M. E. Villa-Pérez, M. Á. Álvarez-Carmona, O. Loyola-González, M. A. Medina-Pérez, J. C. Velazco-Rossell, and K.-K. R. Choo, "Semi-supervised anomaly detection algorithms: A comparative summary and future research directions," *Knowl.-Based Syst.*, vol. 218, p. 106878, Apr. 2021, doi: 10.1016/j.knosys.2021.106878.
- [46] N. Craswell, "Precision at n," in *Encyclopedia of Database Systems*, L. Liu and M. T. Özsu, Eds. New York, NY: Springer, 2016, pp. 1–1. doi: 10.1007/978-1-4899-7993-3\_484-2.
- [47] A. E. Ezugwu et al., "A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects," *Eng. Appl. Artif. Intell.*, vol. 110, p. 104743, Apr. 2022, doi: 10.1016/j.engappai.2022.104743.
- [48] E. H. Budiarto, A. Erna Permanasari, and S. Fauziati, "Unsupervised anomaly detection using K-Means, local outlier factor and one class SVM," in *2019 5th International Conference on Science and Technology (ICST)*, Jul. 2019, vol. 1, pp. 1–5. doi: 10.1109/ICST47872.2019.9166366.
- [49] H. Alimohammadi and S. Nancy Chen, "Performance evaluation of outlier detection techniques in production timeseries: A systematic review and meta-analysis," *Expert Syst. Appl.*, vol. 191, p. 116371, Apr. 2022, doi: 10.1016/j.eswa.2021.116371.
- [50] M. Goldstein, "Anomaly detection in large datasets," Phd thesis (published), Technische Universität Kaiserslautern, Germany, 2014. Accessed: Jul. 24, 2022. [Online]. Available: <https://www.goldiges.de/phd/>.

# Automated Brain Disease Classification using Transfer Learning based Deep Learning Models

Farhana Alam<sup>1</sup>, Farhana Chowdhury Tisha<sup>2</sup>, Sara Anisa Rahman<sup>3</sup>, Samia Sultana<sup>4</sup>  
Md. Ahied Mahi Chowdhury<sup>5</sup>, Ahmed Wasif Reza<sup>6\*</sup>, Mohammad Shamsul Arefin<sup>7</sup>

Department of Computer Science and Engineering, East West University, Dhaka-1212, Bangladesh<sup>1, 2, 3, 4, 5, 6</sup>  
Department of Computer Science and Engineering, Daffodil International University, Dhaka, Bangladesh<sup>7</sup>  
Department of Computer Science and Engineering, Chittagong University of Engineering and Technology<sup>7</sup>  
Chattogram, Bangladesh<sup>7</sup>

**Abstract**—Brain MRI (Magnetic Resonance Imaging) classification is one of the most significant areas of medical imaging. Among different types of procedures, MRI is the most trusted one to detect brain diseases. Manual and semi-automated segmentations need highly experienced radiologists and much time to detect the problem. Recently, deep learning methods have taken attention due to their automation and self-learning techniques. To get a faster result, we have used different algorithms of Convolutional Neural Network (CNN) with the help of transfer learning for classification to detect diseases. This procedure is fully automated, needs less involvement of highly experienced radiologists, and does not take much time to provide the result. We have implemented six deep learning algorithms, which are InceptionV3, ResNet152V2, MobileNetV2, Resnet50, EfficientNetB0, and DenseNet201 on two brain tumor datasets (both individually and manually combined) and one Alzheimer's dataset. Our first brain tumor dataset (total of 7,023 images-training 5,712, testing 1,311) has 99-100 percent training accuracy and 98-99 percent testing accuracy. Our second tumor dataset (total of 3,264 images-training 2,870, testing 394) has 100 percent training accuracy and 69-81 percent testing accuracy. The combined dataset (total of 10,000 images-training 8,000, testing 2,000) has 99-100 percent training accuracy and 98-99 percent testing accuracy. Alzheimer's dataset (total of 6,400 images-training 5,121, testing 1,279, 4 classes of images) has 99-100 percent training accuracy and 71-78 percent testing accuracy. CNN models are renowned for showing the best accuracy in a limited dataset, which we have observed in our models.

**Keywords**—Brain MRI; tumor; deep learning; classification; transfer learning

## I. INTRODUCTION

With the advancement of modern science and technology, brain diseases are still among the deadliest diseases. Magnetic Resonance Imaging (MRI) is a well-known term in the medical sector to diagnose cerebral complications. It is used to detect brain cells that differ from normal cells. There are some other methods such as X-radiation (X-rays), Computed Tomography (CT), Positron Emission Tomography (PET), Single-Photon-Emission Computed Tomography (SPECT), Magnetic Resonance Spectroscopy (MRS), etc. are also used for diagnosis of diseases. But among all of them, MRI is the most popular one to detect problematic cells accurately. MRI is a non-invasive and flexible clinical method that investigates the conditions of the brain and any other body parts in species [1].

\*Corresponding Author.

It uses magnetic fields and radio waves to generate images. For brain MRI, the images are taken from different planes to detect the actual area of the problematic cells of both pre-and post-contrast. Its scanned images provide high contrast and high spatial resolution images, which helps to understand the different characteristics of the soft tissues of a cell. Usually, brain abnormalities are easily found by MRI scans. After analyzing those images, medical experts can easily identify brain disorders such as Alzheimer's disease, schizophrenia, multiple sclerosis, brain tumors, cancer, and degenerative diseases [1]. Although, many neurological diseases need frequent analysis of the brain, in those cases MRI scan is essential.

In the past, segmentation done by humans was a time-consuming procedure and could not provide significant results [2]. On the other hand, automatic segmentation methods result in efficient and precise segmentation. Lately, deep learning methods have been given increasing attention due to their automation and self-learning techniques. Convolutional Neural Network (CNN) is one of the most popular deep learning architectures and it has shown outstanding impact on various industries, such as medical, electronics, robotics, etc. The main advantage of CNN is that it can learn abstract features of the image without having preceding acknowledgment compared to classical methods. This method is developing daily and has achieved numerous appreciations in brain segmentation and classification. Precise segmentation of a 2D and 3D image has always been a challenging task, and various approaches have been proposed for better accuracy in the past. But state-of-the-art deep learning architectures for image segmentation have managed to compute complex 3D models. For these reasons, automatic detection and classification are highly demanding attributes in the decision-making of medical science. Again, CNN models show high accuracy even in limited datasets which are also one of the reasons for choosing CNN models. As we have used transfer learning, the process has become faster. Most of the traditional supervised learning algorithms are not supportive of multi-class classifications as well as very few experiments have been done on recently developed deep learning algorithms for brain MRI classification. Therefore, the question remains what is an efficient way to classify brain diseases from MR images? Also, a comparison of different deep learning algorithms on different types of datasets is



missing, which raises the question of how well a model works on different types of images.

The main objective of this study is to find an efficient way to classify diseases from brain MRI using deep learning models and show a comparative study of them for multi-class brain MRI classification problems. Six CNN models which are commonly used in classification of brain MRI- InceptionV3 [3], ResNet152V2, MobileNetV2 [4], Resnet50 [5], EfficientNetB0 [6] and DenseNet201 [7][8][9][10][11][12][13][14][15][16][17][18][19]. We have implemented these models on three different datasets- one is an Alzheimer's dataset, and the others are brain tumor datasets, all of which are open-access datasets. This study contributes to the health sector, where it is crucial to act in a short time in case of any emergency. Our study can reduce the time to classify a disease from an MR image, which can also lower the occurrence of human error. Also, it can reduce the cost for the patients. As CNN models are getting developed day by day, we could improve our health sector services by finding the most efficient one.

The rest of the paper is organized as follows: In Section II, we have reviewed the related paper materials and their research analogy. A brief overview of publicly available brain MRI datasets, followed by a brain MRI analysis and overview of CNN architectures are discussed in Section III. We have analyzed the performance of our proposed architecture on three publicly available datasets and compared their performance with other methods in Section IV. In Section V, we conclude the paper.

## II. RELATED WORKS

Deep learning models are very recent but many research works have been done for the classification of brain tissues. A method for binary classification of brain tumors is proposed, where they took only the region of interest from MRI images by using Open source Computer Vision (CV) Canny Edge Detection technique and trained a CNN model of eight convolutional layers [20]. Multiclass classification of brain tumors is proposed by selecting features using Densenet201 Pre-Trained Deep Learning Model, Entropy-Kurtosis-based High Feature Values (EKbHFV), and a modified genetic algorithm (MGA), where Cubic SVM classifier is used to classify the selected features after fusing using a non-redundancy-based fusion approach [21]. Again, a CNN model of 18 layers is used for cropped lesions, uncropped lesions, and segmented lesion images for multiclass classification of brain tumors [22].

Some works have either pre-trained data or implemented a single model. MobileNetV2 is used to classify brain tumors with the accuracy of 94% [15], applied ResNet152V2 for classifying four types of brain tumors by using various pre-processing steps to achieve an accuracy of 98.9% [11], and used 29 different pre-trained models to classify Alzheimer's disease, achieved the highest accuracy of 92.98% by EfficientNetB0 [19].

In some literature, they have used multiple planes and multiple layers to detect the problem. A multi-pathway CNN architecture is proposed where the input images are processed

in three spatial scales: sagittal, coronal, and axial views [23] and implemented CNN model with small kernels and neuron weight to classify between tumor and non-tumor which brought 97.5% accuracy with very low complexity [24]. Some works have been done by summing up a few models. Using a method where pre-trained models are used for feature concatenation, it is found that features from the pre-trained model of InceptionV3 and DensNet201 can classify three-class brain tumor datasets better than existing state-of-the-art deep learning methods [8]. Using five CNN models, they used the weighted average of those models to get an accuracy of the 96% in classifying stages of Alzheimer's disease [12]. By removing the last five layers of ResNet50 and adding 8 new layers, achieved 97.2% accuracy in classifying brain tumors and also used Alexnet, Resnet50, Densenet201, InceptionV3, and Googlenet models to classify brain tumors [16].

In some, data are pre-processed in different ways, then models are applied to them. An automated brain disease classification model is created with four main phases, which are preprocessing, exemplar deep feature generator, feature selection, and classification using a support vector machine (SVM) [13]. By using Discrete Cosine Transform-based image fusion, which is combined with a super-resolution and classifier framework, a CNN model, ResNet50, achieved a 98.14% accuracy rate on an open-access dataset [18].

The literature demonstrates a handful of models which have the potential to be used in the brain disease classification sector. However, an in-depth analysis of the most popular and most efficient models on different types of datasets is not quite present in the explored studies.

In our experiment, we have implemented six different models- InceptionV3, ResNet152V2, MobileNetV2, Resnet50, EfficientNetB0, and DenseNet201 on two brain tumor datasets (both individually and manually combined) and an Alzheimer's dataset to visualize the difference of each model, compare their effectiveness by using four different measurements- Accuracy, Precision, Recall, and F1-score, and efficiency.

## III. MATERIALS AND METHODS

### A. Dataset

There is a total of three datasets from three different sources, each containing brain MR images of four kinds. Two datasets contain three variants of Brain Tumor, while one dataset contains three variants of Alzheimer's. We have combined two datasets of Brain Tumor to expand the size of the data and reduce biasness. For simplicity purposes, we have named the datasets respectively D1, D2, D3, and D4. D1 represents the Brain Tumor MRI dataset, D2 as Brain Tumor Classification (MRI), D3 as the manual combined dataset, and finally D4 as the Alzheimer's Dataset (4 classes of image). The datasets are publicly available and collected from Kaggle.

In Fig. 1, Fig. 2, and Fig. 3, glioma, meningioma, pituitary, and no tumor are the variations of brain tumor datasets and in Fig. 4, moderate demented, mild demented, non-demented and very mild demented are the four different classes of Alzheimer's dataset. The summary of all datasets is given in Table I.

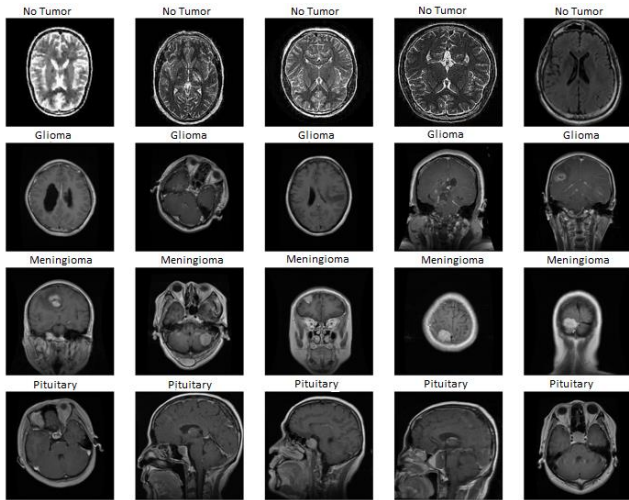


Fig. 1. Sample Images from D1.

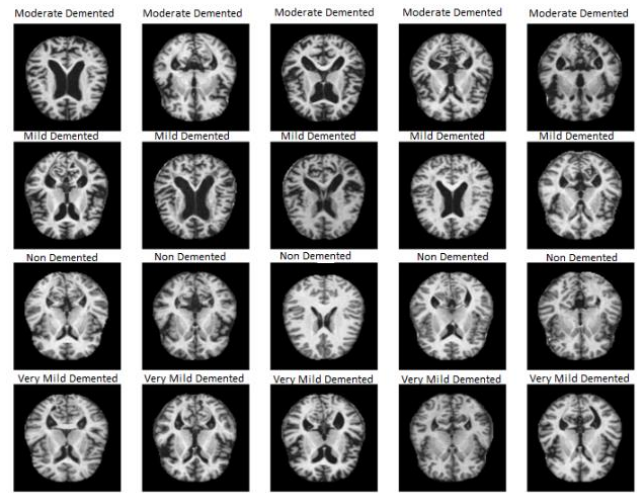


Fig. 4. Sample Images from D4.

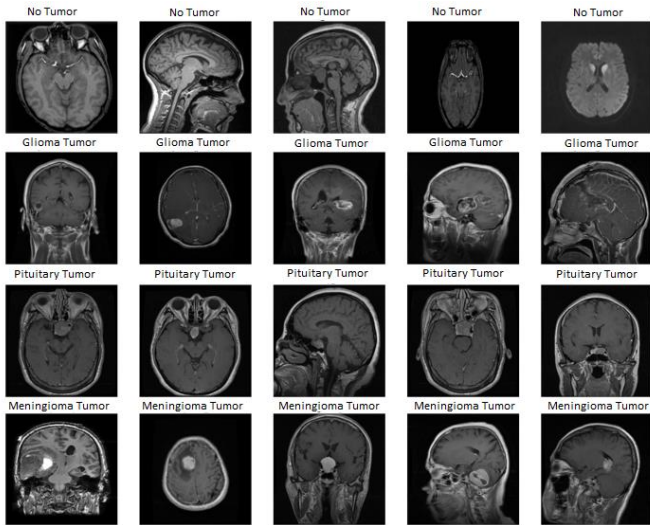


Fig. 2. Sample Images from D2.

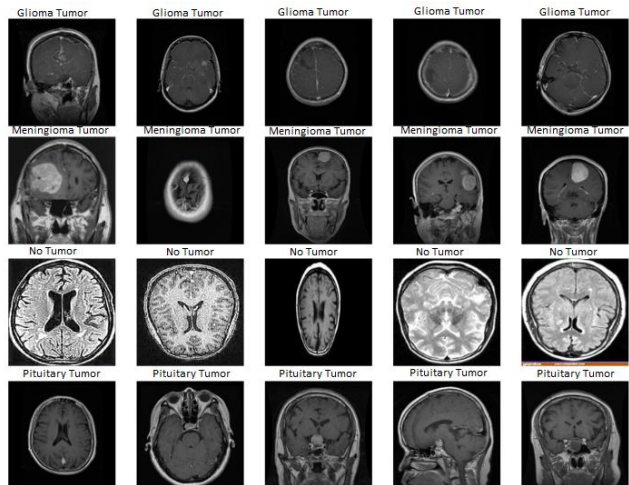


Fig. 3. Sample Images from D3.

TABLE I. SUMMARY OF DATASETS

Dataset	Classes	No. of Images	Total Images
Brain Tumor MRI Dataset (D1)	No Tumor	2000	7023
	Glioma	1621	
	Meningioma	1645	
	Pituitary	1757	
Brain Tumor Classification MRI (D2)	No Tumor	500	3264
	Glioma	926	
	Meningioma	937	
	Pituitary	901	
Manually Combined Dataset (D3)	No Tumor	2500	10000
	Glioma	2500	
	Meningioma	2500	
	Pituitary	2500	
Alzheimer's Dataset (D4)	Moderate Demented	64	6400
	Mild Demented	896	
	Non-Demented	3200	
	Very Mild Demented	2240	

### B. Preprocessing

At first, the images are converted into NumPy arrays, where each pixel of an image is assigned to a number, generating an array for each image. Image Augmentation is an important step in image processing. It creates multiple versions of one image to increase the size of the dataset. Each version has different properties that give more information and a new point of view of the image to train. The images are shifted both vertically and horizontally between -20px to 20px, randomly zoomed in and out by 20%. Then the categorical values are converted to numeric values. As there are four classes for each dataset, each class is given a numeric value in the range 0-3.

### C. Transfer Learning

Transfer learning is a machine learning technique where a deep learning model reuses the weights that have been generated from a different dataset. The reason for using this method is to use the patterns learned from a similar task to get a head start to avoid huge computational time and attain the best result possible for that model.

The models used in this experiment have already been trained with the ImageNet dataset, which provided us with the weights that we can utilize.

### D. Models

Six deep learning models are used train the datasets, these are InceptionV3, ResNet152V2, MobileNetV2, ResNet50, EfficientNetB0, and DenseNet201.

1) *InceptionV3*: InceptionV3 is the third version of Google's Deep Learning Convolutional Architectures series, Inception. It contains 42-48 layers, which include convolutions, max pooling, average pooling, dropouts, and fully connected layers, and has both symmetric and asymmetric building blocks. This model estimates the marginalized effect of label dropout during training to regularize the classifier layer by changing the label-smoothing regularization (LSR) which is defined by,

$$q'(k) = (1 - \epsilon)\delta_{k,y} + \frac{\epsilon}{k} \quad (1)$$

where the uniform distribution  $u(k) = \frac{1}{k}$  is used in the model.

Also by considering the cross entropy, LSR is

$$H(q', p) = (1 - \epsilon)H(q, p) + \epsilon H(u, p) \quad (2)$$

LSR prevents the largest logit or unnormalized log probabilities from becoming much larger than all others. It encourages the model to be less confident as it might cause over-fitting and reduce the adapting capability of the model. InceptionV3 gave more than 78.1% accuracy on the ImageNet Dataset [3].

2) *MobileNetV2*: MobileNetV2 has 53 layers, one average pool, and around 350 GFLOPs (Floating point operations per second). It has two types of convolutional layers: 1x1 Convolution and 3x3 Depthwise Convolution. It contains two main blocks, the Inverted Residual Bottleneck Block and Bottleneck Residual Block.

The inverted residual bottleneck layers are implemented in a memory-efficient way to it can be used for mobile applications. It builds a directed acyclic compute hypergraph G, where the edges are the operations and nodes are tensors of intermediate computation. The target is to minimize the total number of tensors stored in memory, so it selects the computation order  $\Sigma(G)$  which has the minimum memory,

$$M(G) = \min_{\pi \in \Sigma(G)} \max_{i \in 1..n} [\sum_{A \in R(i, \pi, G)} |A|] + size(\pi_i) \quad (3)$$

As for graphs with only trivial parallel structure, the memory needed to compute graph G is

$$\max_{op \in G} \left[ \sum_{A \in op_{inp}} |A| + \sum_{B \in op_{out}} |B| + |op| \right] \quad (4)$$

In the bottleneck residual block, a bottleneck block operator  $F(x)$  can be represented as  $F(x) = [A \circ N \circ B]x$ . As the chain of t tensors of size n/t are the inner tensor, the function can be written as

$$F(x) = \sum_{i=1}^t (A_i \circ N \circ B_i)(x) \quad (5)$$

When t is a small constant between two and five, this method is the most helpful as it can reduce the memory, but can still utilize most of the efficiencies of highly optimized deep learning frameworks [4].

3) *ResNet50*: ResNet50 is a variant of the Residual Network model. It contains 48 convolutional layers, 1 Max Pool, and 1 Average Pool layer with  $3.8 \times 10^9$  floating point operations per second (FLOPs).

The convolutional layers have 3x3 filters. The layers have the same amount of filters when the output feature map size is the same. However, the layers have the double amount of filters when the feature map size is half. By following these two rules, it performs downsampling. The final layer contains an average pool with a fully-connected layer of 1000 nodes with a softmax function. The total number of weighted layers is 34. Based on this network, shortcut connections are inserted which turn this network into its counterpart residual version. The identity shortcut is defined by

$$y = F(x, \{W_i\}) + x \quad (6)$$

Equation (6) can be directly used when the dimension of x and F are equal. But if the dimensions change, either identity mapping is still performed by the shortcut, or the projection shortcut is used to match dimensions, defined by

$$y = F(x, \{W_i\}) + W_s x \quad (7)$$

For both options, when the shortcuts go across feature maps of two sizes, they are performed with a stride of two. By replacing each two-layer block in the 34-layer net with a three-layer bottleneck block, it results in a 50-layer ResNet [5].

4) *EfficientNetB0*: EfficientNetB0 is the base model of EfficientNet family. It uses Model Scaling, where the existing model is scaled based on model width, depth, and resolution. This model introduces a new compound scaling method that uniformly scales the width, depth, and resolution to achieve better accuracy using a compound coefficient  $\emptyset$ . By considering depth,  $d = \alpha^\emptyset$ , width,  $w = \beta^\emptyset$ , and resolution,  $r = \gamma^\emptyset$ , this method is defined as

$$\alpha \cdot \beta^2 \cdot \gamma^2 \approx 2, \text{ where } \alpha \geq 1, \beta \geq 1, \gamma \geq 1 \quad (8)$$

For this principle in (8), for any new  $\emptyset$ , the total FLOPs will increase by  $2^\emptyset$  approximately. EfficientNetB0 is scaled up by using the compound scaling method where  $\emptyset$  is fixed at 1, and after doing a small grid search of  $\alpha, \beta$ , and  $\gamma$ , the best values are found to be 1.2, 1.1, and 1.15 respectively under the constraint of (8).

The model has a total of 237 layers. It consists of five different modules which are used in a certain way to create each block of the model [6].

5) *DenseNet201*: DenseNet201 is one of the models of the DenseNet group. It contains 201 layers, and it is divided into Dense Blocks with different filters and the same dimensions for each block. The network includes  $L(L+1)/2$  direct connections. The output of the previous layer becomes the input of the next layer by using composite function operations. The transition layer is added between the Dense Blocks, where it applies batch normalization using downsampling. The growth rate  $k$  controls how much information should be added to the next layer. At layer  $l$  the growth rate is defined by [7]:

$$k^{[l]} = \left( k^{[0]} + k(l - 1) \right) \quad (9)$$

The models are compiled for training by using the compile method of Keras Model Training API. The training data is fit into the model with 10% validation data, 15 epochs, and a batch size of 32. The categorical accuracy of each epoch is monitored to find the highest categorical accuracy by following (10).

$$\text{Categorical Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}} \quad (10)$$

#### IV. RESULTS AND DISCUSSION

##### A. Result Analysis

We have implemented six algorithms on four different datasets and secured significant results. The maximum accuracy for each dataset was obtained by different models. As we have used various datasets, so the percentage of the accuracy was not consistent in every dataset.

Our models were trained by using 80 percent of the data for training, 10 percent of the training data for validation, and 20 percent for testing. We have used ImageNet as a pre-trained weight in every model and SoftMax for the pre-training classifier. We also included some hyper-parameters as well.

The following figures show how the training and validation accuracy/loss fluctuated per epoch in four different datasets for the models which gave the best results. The red line represents the validation accuracy/loss and the green one is for training accuracy/loss. Also, the final results of each of the models from different datasets are given in Tables II, III, IV, and V. The model with the best accuracy is highlighted in the tables.

TABLE II. ACCURACY SUMMERY OF D1

Model	Training Accuracy	Validation Accuracy	Testing Accuracy
<b>InceptionV3</b>	<b>100%</b>	<b>97.90%</b>	<b>99.54%</b>
ResNet152V2	100%	96.50%	98.63%
MobileNetV2	99.98%	94.58%	98.09%
ResNet50	100%	98.78%	98.63%
EfficientNetB0	99.98%	99.65%	99.47%
DenseNet201	100%	96.85%	99%

TABLE III. ACCURACY SUMMERY OF D2

Model	Training Accuracy	Validation Accuracy	Testing Accuracy
InceptionV3	100%	98.26%	76.65%
ResNet152V2	100%	100%	76.90%
MobileNetV2	100%	92.68%	69.04%
ResNet50	100%	99.30%	77.16%
<b>EfficientNetB0</b>	<b>100%</b>	<b>96.17%</b>	<b>81.47%</b>
DenseNet201	100%	97.56%	78.43%

TABLE IV. ACCURACY SUMMERY OF D3

Model	Training Accuracy	Validation Accuracy	Testing Accuracy
InceptionV3	100%	98.50%	93.45%
ResNet152V2	100%	98.38%	93.10%
<b>MobileNetV2</b>	<b>99.94%</b>	<b>98.75%</b>	<b>94.50%</b>
ResNet50	100%	98.13%	93.45%
EfficientNetB0	100%	99.38%	94.35%
DenseNet201	100%	98.75%	94.10%

TABLE V. ACCURACY SUMMERY OF D4

Model	Training Accuracy	Validation Accuracy	Testing Accuracy
InceptionV3	100%	99.42%	77.25%
ResNet152V2	100%	98.64%	71.07%
MobileNetV2	100%	100%	76.15%
ResNet50	100%	99.42%	73.50%
<b>EfficientNetB0</b>	<b>99.98%</b>	<b>99.03%</b>	<b>78.34%</b>
DenseNet201	100%	99.61%	75.45%

As we can see different model provides different accuracy based on different datasets. In the dataset D1 shown in Table II, we have achieved 99.54 percent testing accuracy using the InceptionV3 model and the minimum was 98.09 percent using MobileNetV2. In the D2 dataset shown in Table III, we have secured 81.47 percent testing accuracy using the EfficientNetB0 model and 69.04 percent was the minimum by MobileNetV2. After analyzing the second dataset we noticed that there was no equal distribution among the classes, which is why the accuracy might not meet its target. Therefore, we have combined D1 and D2 by maintaining the equal distribution of the four classes and have generated D3. D3 has received improved testing accuracy which is 94.50 percent using MobileNetV2 and ResNet152V2 provided its minimal 93.10 percent shown in Table IV. We have also implemented these models on a different dataset which consisted of the images of Alzheimer's disease and the maximum accuracy has been received at 78.34 percent by EfficientNetB0 shown in Table V. So, we can conclude that InceptionV3, EfficientNetB0, and MobileNetV2 these three models are working best CNN models so far according to our observation.

ResNet models are found to perform comparatively worse than other models. It focuses mainly on creating a deep neural network model without hampering the accuracy. As a result, it takes longer as it has a relatively deep architecture, i.e. more parameters to train. Furthermore, having a deep architecture can also be the reason for its consistent validation loss. In contrast, as we can see in Fig. 5 and Fig. 7 that InceptionV3 and MobileNetV2 take significantly less time for performance. They have divided the convolution layer into two distinct parts. Firstly, instead of separately applying the kernel to all the channels, they have applied depthwise convolution. It mainly applies the kernel to each of the channels individually. Secondly, they applied a convolution with one kernel size to combine the features of the newly generated channel, which directly contributes to the shorter training period. However, as it is apparent from our results, it also sacrifices the overall accuracy. In contrast to MobileNetV2 and InceptionV3, EfficientNet evaluates the scaling part of the neural network. Using a compound coefficient, they uniformly scale the width, resolution, and depth of the network simultaneously to find the best gains. As it is apparent from our results shown in Fig. 6 and Fig. 8, EfficientNetB0 provides significantly better accuracy while taking comparatively less training time. Table VI shows the training and prediction time taken by each model.

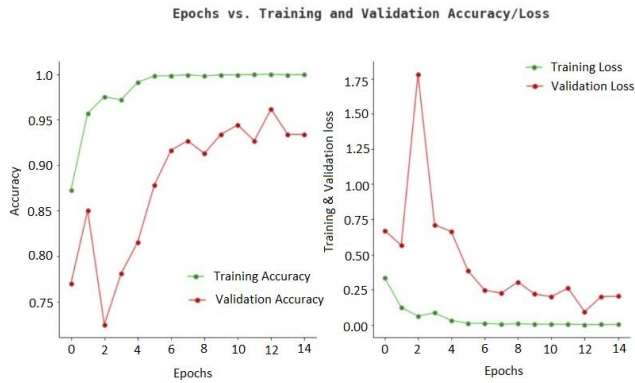


Fig. 5. Epochs vs. Training and Validation Accuracy/Loss of D2 (EfficientNetB0).

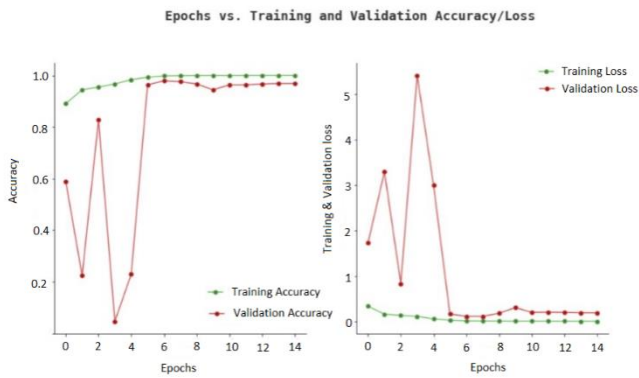


Fig. 6. Epochs vs. Training and Validation Accuracy/Loss of D1 (InceptionV3).

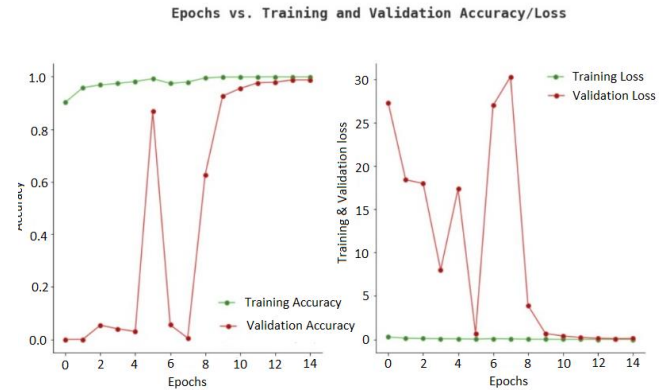


Fig. 7. Epochs vs. Training And Validation Accuracy/Loss of D3 (MobileNetV2).

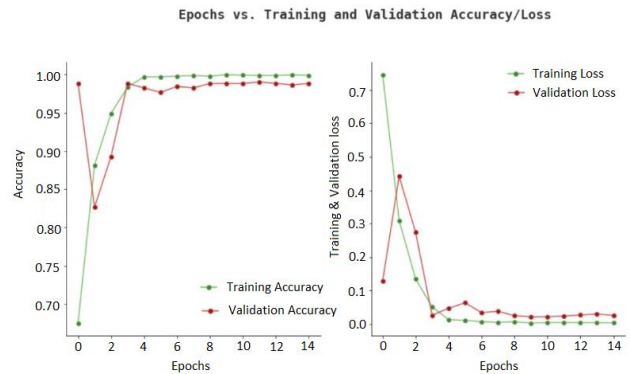


Fig. 8. Epochs vs. Training and Validation Accuracy/Loss of D4 (EfficientNetB0).

TABLE VI. TRAINING AND PREDICTION TIME COMPARISONS

Dataset	Model	Training Time (sec)	Prediction Time (sec)
D1	InceptionV3	~ 1480	~ 8.5
	ResNet152V2	~ 4033 (Worst Case)	~ 12.8
	MobileNetV2	~ 895	~ 4.9
	ResNet50	~ 1706	~ 8.8
	EfficientNetB0	~ 764 (Best Case)	~ 4.2
	DenseNet201	~ 1387	~ 5.6
D2	InceptionV3	~ 778	~ 2.8
	ResNet152V2	~ 925 (Worst Case)	~ 3.6
	MobileNetV2	~ 264 (Best Case)	~ 1.4
	ResNet50	~ 428	~ 4.7
	EfficientNetB0	~ 388	~ 3.1
	DenseNet201	~ 748	~ 2.6
D3	InceptionV3	~ 2573	~ 13.3
	ResNet152V2	~ 5426	~ 15.1
	MobileNetV2	~ 17, 282 (Worst Case)	~ 23.2
	ResNet50	~ 2275	~ 9.8
	EfficientNetB0	~ 1812 (Best Case)	~ 9.4
	DenseNet201	~ 4020	~ 15.7
D4	InceptionV3	~ 591	~ 5.3
	ResNet152V2	~ 1670 (Worst Case)	~ 10.6
	MobileNetV2	~ 457 (Best Case)	~ 2.6
	ResNet50	~ 703	~ 4.6
	EfficientNetB0	~ 670	~ 3.8
	DenseNet201	~ 1276	~ 6.3



As we know validation of these models is very important to prove our performance for this research purpose, we have used some metrics to do so- precision, recall, f1 score, and confusion matrix. The mathematical notations of these terms are given below:

$$Precision = \frac{true\ positive}{true\ positive + false\ positive} \quad (11)$$

$$Recall = \frac{true\ positive}{true\ positive + false\ negative} \quad (12)$$

$$F1\ Score = 2 \times \frac{precision \times recall}{precision + recall} \quad (13)$$

The confusion matrices in Fig. 9 to 12, and the evaluation results calculated by (11), (12), and (13) in Tables VII, VIII, IX, and X of the models with the best results from each dataset are shown below:

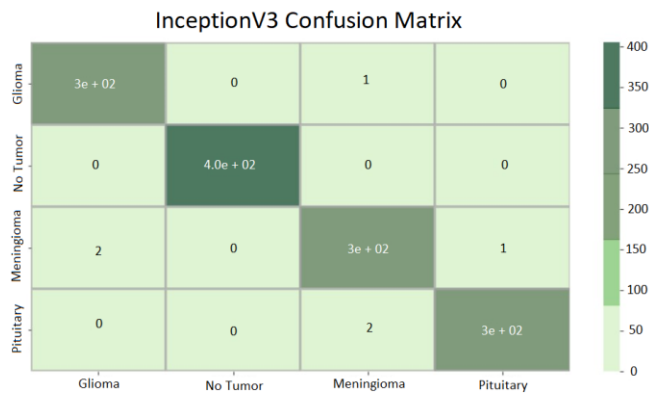


Fig. 9. Confusion Matrix of D1.

TABLE VII. EVALUATION RESULT OF D1 (INCEPTIONV3)

Class	Precision	Recall	F1-Score	Accuracy
Glioma	0.99	1.00	1.00	1.00
Meningioma	1.00	1.00	1.00	
Pituitary	0.99	0.99	0.99	
No Tumor	1.00	0.99	0.99	

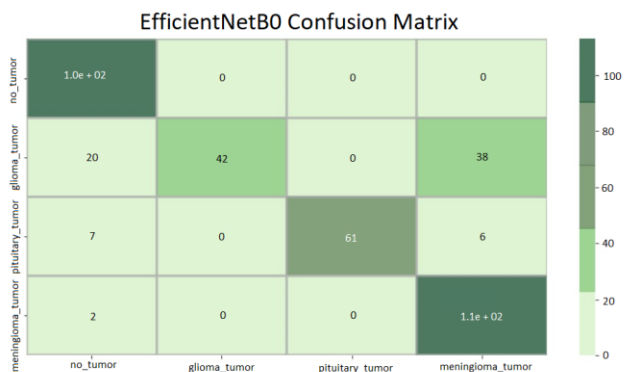


Fig. 10. Confusion Matrix of D2.

TABLE VIII. EVALUATION RESULT OF D2 (EFFICIENTNETB0)

Class	Precision	Recall	F1-Score	Accuracy
Glioma	0.78	1.00	0.88	0.81
Meningioma	1.00	0.42	0.59	
Pituitary	1.00	0.82	0.90	
No Tumor	0.72	0.98	0.83	

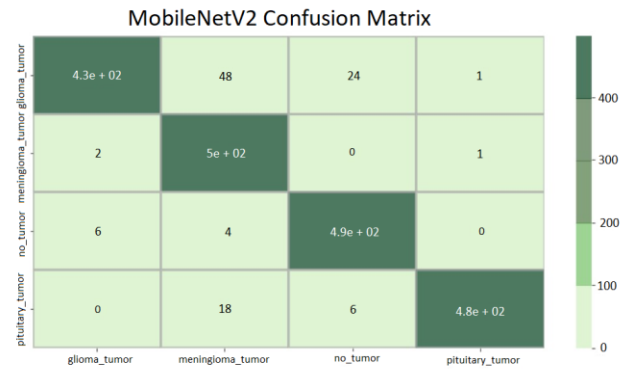


Fig. 11. Confusion Matrix of D3.

TABLE IX. EVALUATION RESULT OF D3 (MOBILENETV2)

Class	Precision	Recall	F1-Score	Accuracy
Glioma	0.98	0.85	0.91	0.94
Meningioma	0.88	0.99	0.93	
Pituitary	0.94	0.98	0.96	
No Tumor	1.00	0.95	0.97	

EfficientNetB0 Confusion Matrix

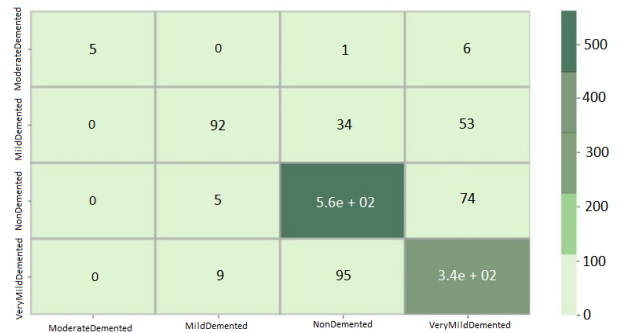


Fig. 12. Confusion Matrix of D4.

TABLE X. EVALUATION RESULT OF D4 (EFFICIENTNETB0)

Class	Precision	Recall	F1-Score	Accuracy
Moderate Demented	1.00	0.42	0.59	0.78
Mild Demented	0.87	0.51	0.65	
Non Demented	0.81	0.88	0.84	
Very Mild Demented	0.72	0.77	0.74	



### B. Performance

For comparison purposes, we have mentioned different types of models and their accuracy percentages in the relevant sector.

TABLE XI. COMPARISON TABLE WITH OTHER MODELS

Serial	Model	Dataset Size	Accuracy (%)
1.	DCT-CNN-ResNet50 [18]	70,220	98.14%
2.	ELM-LRF [25]	220,875	97.18%
3.	Multiscale Convolutional Neural Network [23]	3,264	97.30%
4.	Multiclass SVM cubic classifier [21]	335	99.8%
5.	InceptionV3 [10]	411	85%
6.	ResNet152V2 [11]	7,023	98.90%
7.	MobileNetV2 [15]	2,475	94%
8.	EfficientNetB0 [19]	6,400	92.98%
9.	InceptionV3 (Our model)	7,023	99.54%
10.	EfficientNetB0 (Our model)	3,264	81.47%
11.	MobileNetV2 (Our model)	10,000	94.50%
12.	EfficientNetB0 (Our model)	6,400	78.34%

From Table XI, we can see many other studies have used CNN models as well. In [11], the author found the highest accuracy of 98.90% by the ResNet152V2 model, where we have managed to use InceptionV3 to achieve 99.54% using the same dataset. However, our EfficientNetB0 model did not outperform the model used in [23] and [19], where both datasets were identical. Although for other models all of us did not use the same dataset, we cannot compare the accuracy for every model (e.g., [26]) entirely.

### C. Challenges

The study uses deep learning models, which is a very time-consuming procedure even with the help of transfer learning. The amount of data was not enough, generating some difficulties while getting a better result. Also, the use of medical datasets introduced its own challenges, because this sector has an extreme restriction on the time limit and the results have to be monitored as accurately as possible, as the application of these models in actual medical settings is expected in the future.

### V. CONCLUSION

Our goal was to find which CNN model or models can provide the maximum result. That is why we have implemented six models in four datasets of different types. Our datasets were not large enough to study in this field and not all datasets are suitable for every algorithm, which resulted in different accuracy in different datasets with the same algorithm. As we are using medical data, it is challenging to predict exact results, yet we have secured a remarkable accuracy of 99.54 percent. In medical science, time is one of the most crucial factors for any emergency. Therefore, by implementing the most promising CNN models, we can be able to analyze brain MRI images straight away. Mostly in our

country, doctors try to identify the types of tumors manually as well as any other brain diseases, so the risk of the occurrence of human error is high. Sometimes patients would have to use high-priced diagnostic methods to find the types of the disease. Health-related issues are way too sensitive, hence any type of mistake is unacceptable. Therefore, this study could save time and cost, and most importantly could save lives.

By using the most prominent model, we can develop software that can be used to detect a tumor or any other disease in an instant. We have only worked with brain MRI images, in the future, we would like to work with different types of MRI images as well.

### ACKNOWLEDGMENT

We are grateful to Dr. Bidyut Kumar Saha, MBBS, MD (India), Post Doctoral Training in Neuro Cardio Radiology (AIIMS, Delhi), Senior Consultant & Coordinator – Diagnostic & Interventional Radiology, Asgar Ali Hospital. He guided and collaborated with us and shared his medical knowledge regarding our dataset analyzing part. Without his help, we would not be able to do our tasks flawlessly.

### REFERENCES

- [1] Z. Akkus, A. Galimzianova, A. Hoogi, D. L. Rubin, and B. J. Erickson, "Deep Learning for Brain MRI Segmentation: State of the Art and Future Directions," *J. Digit. Imaging*, vol. 30, no. 4, pp. 449–459, 2017, doi: 10.1007/s10278-017-9983-4.
- [2] I. Despotović, B. Goossens, and W. Philips, "MRI segmentation of the human brain: Challenges, methods, and applications," *Comput. Math. Methods Med.*, vol. 2015, 2015, doi: 10.1155/2015/450341.
- [3] C. Szegedy, V. Vanhoucke, S. Ioffe, and J. Shlens, "Rethinking the Inception Architecture for Computer Vision."
- [4] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. C. Chen, "MobileNetV2: Inverted Residuals and Linear Bottlenecks," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 4510–4520, 2018, doi: 10.1109/CVPR.2018.00474.
- [5] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," Accessed: Feb. 05, 2022. [Online]. Available: <http://image-net.org/challenges/LSVRC/2015/>.
- [6] M. Tan and Q. V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks."
- [7] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," Accessed: Feb. 05, 2022. [Online]. Available: <https://github.com/liuzhuang13/DenseNet>.
- [8] N. Noreen, S. Palaniappan, A. Qayyum, I. Ahmad, M. Imran, and M. Shoaib, "A Deep Learning Model Based on Concatenation Approach for the Diagnosis of Brain Tumor," *IEEE Access*, vol. 8, pp. 55135–55144, 2020, doi: 10.1109/ACCESS.2020.2978629.
- [9] S. S. Yadav and S. M. Jadhav, "Deep convolutional neural network based medical image classification for disease diagnosis," *J. Big Data*, vol. 6, no. 1, pp. 1–18, Dec. 2019, doi: 10.1186/S40537-019-0276-2/TABLES/16.
- [10] A. Kumar, P. Pathak, and P. Stynes, "A Transfer Learning Approach to Classify the Brain Age from MRI Images," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 12581 LNCS, pp. 103–112, Dec. 2020, doi: 10.1007/978-3-030-66665-1\_8.
- [11] A. Alnemer and J. Rasheed, "An Efficient Transfer Learning-based Model for Classification of Brain Tumor," *ISMSIT 2021 - 5th Int. Symp. Multidiscip. Stud. Innov. Technol. Proc.*, pp. 478–482, 2021, doi: 10.1109/ISMSIT52890.2021.9604677.
- [12] S. U. Sadat, H. H. Shomee, A. Awwal, S. N. Amin, M. T. Reza, and M. Z. Parvez, "Alzheimer's Disease Detection and Classification using Transfer Learning Technique and Ensemble on Convolutional Neural

- Networks,” pp. 1478–1481, Jan. 2022, doi: 10.1109/SMC52423.2021.9659179.
- [13] A. Kursad Poyraz, S. Dogan, E. Akbal, and T. Tuncer, “Automated brain disease classification using exemplar deep features,” *Biomed. Signal Process. Control*, vol. 73, p. 103448, Mar. 2022, doi: 10.1016/J.BSPC.2021.103448.
- [14] Y. K. Cetinoglu, I. O. Koska, M. E. Uluc, and M. F. Gelal, “Detection and vascular territorial classification of stroke on diffusion-weighted MRI by deep learning,” *Eur. J. Radiol.*, vol. 145, p. 110050, Dec. 2021, doi: 10.1016/J.EJRAD.2021.110050.
- [15] T. H. Arfan, M. Hayaty, and A. Hadinegoro, “Classification of Brain Tumours Types Based on MRI Images Using Mobilenet,” *2021 2nd Int. Conf. Innov. Creat. Inf. Technol. ICITech 2021*, pp. 69–73, Sep. 2021, doi: 10.1109/ICITECH50181.2021.9590183.
- [16] A. Çinar and M. Yildirim, “Detection of tumors on brain MRI images using the hybrid convolutional neural network architecture,” *Med. Hypotheses*, vol. 139, p. 109684, Jun. 2020, doi: 10.1016/J.MEHY.2020.109684.
- [17] L. V. Fulton, D. Dolezel, J. Harrop, Y. Yan, and C. P. Fulton, “Classification of Alzheimer’s Disease with and without Imagery Using Gradient Boosted Machines and ResNet-50,” *Brain Sci.* 2019, Vol. 9, Page 212, vol. 9, no. 9, p. 212, Aug. 2019, doi: 10.3390/BRAINSCI9090212.
- [18] A. Deshpande, V. V. Estrela, and P. Patavardhan, “The DCT-CNN-ResNet50 architecture to classify brain tumors with super-resolution, convolutional neural network, and the ResNet50,” *Neurosci. Informatics*, vol. 1, no. 4, p. 100013, Dec. 2021, doi: 10.1016/J.NEURI.2021.100013.
- [19] S. Savaş, “Detecting the Stages of Alzheimer’s Disease with Pre-trained Deep Learning Architectures,” *Arab. J. Sci. Eng.* 2021, pp. 1–18, Sep. 2021, doi: 10.1007/S13369-021-06131-3.
- [20] H. A. Khan, W. Jue, M. Mushtaq, and M. U. Mushtaq, “Brain tumor classification in MRI image using convolutional neural network,” *Math. Biosci. Eng.*, vol. 17, no. 5, pp. 6203–6216, 2020, doi: 10.3934/MBE.2020328.
- [21] M. I. Sharif, M. A. Khan, M. Alhussein, K. Aurangzeb, and M. Raza, “A decision support system for multimodal brain tumor classification using deep learning,” *Complex Intell. Syst.*, no. 0123456789, 2021, doi: 10.1007/s40747-021-00321-0.
- [22] A. M. Alqudah, H. Alquraan, I. A. Qasmieh, A. Alqudah, and W. Al-Sharu, “Brain tumor classification using deep learning technique - A comparison between cropped, uncropped, and segmented lesion images with different sizes,” *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 8, no. 6, pp. 3684–3691, 2019, doi: 10.30534/ijatcse/2019/155862019.
- [23] F. J. Díaz-Pernas, M. Martínez-Zarzuola, D. González-Ortega, and M. Antón-Rodríguez, “A deep learning approach for brain tumor classification and segmentation using a multiscale convolutional neural network,” *Healthc.*, vol. 9, no. 2, 2021, doi: 10.3390/healthcare9020153.
- [24] J. Seetha and S. S. Raja, “Brain tumor classification using Convolutional Neural Networks,” *Biomed. Pharmacol. J.*, vol. 11, no. 3, pp. 1457–1461, 2018, doi: 10.13005/bpj/1511.
- [25] A. Ari and D. Hanbay, “Deep learning based brain tumor classification and detection system,” *Turkish J. Electr. Eng. Comput. Sci.*, vol. 26, no. 5, pp. 2275–2286, 2018, doi: 10.3906/elk-1801-8.
- [26] M. F. Siddiqui, G. Mujtaba, A. W. Reza, and L. Shuib, “Multi-class disease classification in brain MRIs using a computer-aided diagnostic system,” *Symmetry*, vol. 9, no. 3, pp. 1–14, 2017, doi: 10.3390/sym9030037.

# Toward A Holistic, Efficient, Stacking Ensemble Intrusion Detection System using a Real Cloud-based Dataset

Ahmed M. Mahfouz<sup>1</sup>, Abdullah Abuhussein<sup>2</sup>, Faisal S. Alsubaei<sup>3</sup>, Sajjan G. Shiva<sup>4</sup>  
Department of Computer Science, University of Memphis, Memphis, TN 38152, USA<sup>1,4</sup>  
Information Systems Department, St. Cloud State University, St. Cloud, MN 56301, USA<sup>2</sup>  
Department of Cybersecurity, University of Jeddah, Jeddah 23890, Saudi Arabia<sup>3</sup>

**Abstract**—Network intrusion detection is a key step in securing today's constantly developing networks. Various experiments have been put forward to propose new methods for resisting harmful cyber behaviors. Though, as cyber-attacks turn out to be more complex, the present methodologies fail to adequately solve the problem. Thus, network intrusion detection is now a significant decision-making challenge that requires an effective and intelligent approach. Various machine learning algorithms such as decision trees, neural networks, K nearest neighbor, logistic regression, support vector machine, and Naive Bayes have been utilized to detect anomalies in network traffic. However, such algorithms require adequate datasets to train and evaluate anomaly-based network intrusion detection systems. This paper presents a testbed that could be a model for building real-world datasets, as well as a newly generated dataset, derived from real network traffic, for intrusion detection. To utilize this real dataset, the paper also presents an ensemble intrusion detection model using a meta-classification approach enabled by stacked generalization to address the issue of detection accuracy and false alarm rate in intrusion detection systems.

**Keywords**—Intrusion detection system; IDS dataset; stacking ensemble ids; stacking; security; ensemble learning

## I. INTRODUCTION

With the exponential growth of network-based applications globally, there has been a transformation in the business models of organizations [1]. Cost reduction of both computational devices and the Internet have led people to become more technology dependent. As a result of the increasing use of computer networks, new risks have emerged [2]. Therefore, the process of enhancing the speed and precision of security mechanisms has become crucial. Although abundant new security tools have been developed, the rapid evolution of malicious actions continues to be a demanding matter, as their ever-evolving attacks continue to create huge threats to network security [3]. Classical security techniques—for instance, firewalls—are used as a first line of defense against security problems but remain unable to detect internal intrusions or adequately provide security countermeasures [4]. Thus, network administrators tend to rely predominantly on Intrusion Detection Systems (IDSs) to detect such network intrusive actions.

During the past decade, it has become clear that the trend of using the cloud services model in preference to the old on-premises model is increasing rapidly for many reasons [5]. For

instance, the unique utilization/charging models offered by the cloud provider that gives customers the flexibility to adjust their expenses easily, based on their needs. Scaling processes would consume much more time, effort, and expense without the cloud model. With the cloud model, the Capital Expenditure (CapEx) is reduced to the minimum or removed. These elements are taken care of by the cloud provider, which reduces the time to market (TTM) of the services and facilitates hunting market opportunities. With these merits, and many more, adopting the cloud model enables the customer to focus on service development rather than infrastructure management, which helps in achieving customer satisfaction and maximizing revenue. However, using the cloud model comes with many implications and consequences, especially on the security side of the model. One such implication is the huge increase in the number of machines exposed to the Internet since the management of those remote servers, hosted over the cloud, by legitimate users, entails enabling remote access to the servers, which increases the number and kind of vulnerabilities that can be exploited by the attackers.

With the advances in the field of machine learning, studying the malicious traffic patterns and the attacker's behavior for the purpose of developing detection and mitigation/reduction algorithms has become a hot area of research [6]. A vital building block of most of the machine learning techniques is the dataset that is used either in the training phase in case of unsupervised learning or the training and testing phases in case of supervised learning. Due to the significance of the dataset (as shown later), many studies have been devoted to generating such a dataset using different techniques and setups [7].

Most of the time, the datasets used in different studies depend on a simulated dataset due to the lack of publicly available real datasets of the network attacks [8]. This is mainly attributed to the fact that organizations are usually hesitant to publicly share technical information with others about their computing assets, such as applications, network layout, or other information that can be extracted/guessed from a dataset. Doing so risks exposing confidential and sensitive data about the organization's computing assets from security and business perspectives and costs a lot more than taking the risk of sharing. Another reason for the scarcity of real datasets is that they would reveal valuable information about the organization's Intrusion Detection System (IDS) if the machine learning algorithm is trained on the same dataset, and this could help intruders to

bypass it. Although resorting to a simulated dataset seems to be a good solution, it could result in less accurate algorithms when applied in real-world systems [9]. Aside from being simulated or real, attack datasets used for machine learning models have a conceptual problem, which is the imbalance issue since the attacker would be trying to hide his traffic in the normal user traffic. Another shortcoming in the existing datasets is that most of them are a bit outdated, and most of the efforts focus on the attacks, but not the pre- or post-attack (attacker's behavior).

This paper produces a new network intrusion dataset based on real network attacks on up-to-date cloud-based infrastructure. It also offers an adaptive ensemble classifier model, which integrates the advantages of different Machine Learning (ML) classifiers for diverse kinds of attacks and achieves best results using ensemble learning. The proposed model uses a meta-classification method based on stacked generalization for network IDS. The advantage of ensemble learning is combining the predictions of numerous base estimators to expand generalizability and strength over that of a single estimator.

## II. RELATED WORK

### A. Dataset

The effectiveness of any study, or the accuracy of any algorithm that uses a dataset, greatly depends on the dataset quality in terms of both being correctly labelled and being up to date and able to capture the latest attacks [10]. Also, the more data instances there are in the dataset, the greater the accuracy of the experiments and the generalizability of the model. Network attack datasets are constructed by system logs, network logs, network flows, and memory dumps. A novel technique called generative adversarial networks is used to train a generator to create the dataset [11]. The dataset could be built using real or simulated data. The work of one group of researchers [12] provides a comprehensive overview of the existing datasets by analyzing 715 research articles. They focus on three aspects: the origin of the dataset (e.g., real-world vs. synthetic), whether datasets were released by the researchers or not, and the types of datasets that exist. They conclude that 56.4% of the datasets are generated via experiments, while 36.7% are real data. Also, 54.4% of the studies use existing datasets, while the rest created their own, and only 3.8% of them released their datasets. In another research project [13], the authors provide a comprehensive overview of the most used available datasets. Based on their research, the main limitations of the current datasets can be summarized as follows:

- Some of the datasets are old, so they do not help with the recent types of attacks.
- The dataset is not labeled, making it useless for training supervised machine learning models, unless manually labeled, which can be cumbersome.
- The dataset is limited to specific types of attacks or targets specific applications, reducing its generality.
- The dataset is small and does not contain enough data to generalize the trained model.
- The dataset contains redundant data, which could lead to biased models.

- The dataset is completely generated in the lab, making it less representative of the real-world attacks.
- The dataset consists of an imbalanced amount of attack data and benign traffic.

To address the above limitations, one study [14] proposed a dataset approach called CIDD (Cloud Intrusion Detection Dataset) for masquerade attacks. They developed a log analyzer and correlator system to parse and analyze the data from the network. These parsed data are fed to the log analyzer and correlator for processing and marking. The analyzer correlates the user audits in network and host environments using user IP and audit time. Then it assigns user audits to a set of VMs (Virtual Machines) according to their login sessions time and the characteristic of the user task. Finally, it uses the attack and masquerade tables provided by the MIT group to mark the malicious records. The drawback of this dataset is that it lacks the representation of real network traffic as well as actual attack simulations. Moreover, it is outdated for the adequate evaluation of modern IDSs on current networks, regarding types of attacks and the network infrastructure.

In other research [15], the researchers developed a testbed to generate their dataset. The testbed is composed of different machines in a Windows domain and each machine has different types of agents to collect logs and send them to the logger. These machines also have scripts to enable the simulation of some types of attacks, pushed by the logger server, as well as the generation of the normal traffic. The logger server is equipped with the necessary applications to play different roles. Examples of these are an elastic search to collect logs from the whole system, a Mitre Caldera Server to simulate various types of attacks using the installed agents on the hosts, an IDS Suricata for identifying network attack signatures in traffic that is used for labeling the dataset, and others. Unfortunately, the proposed testbed also does not represent real-world network traffic and lacks the actual attacks representation.

### B. Stacking Ensemble IDS

Ensemble learning based methods apply collections of ML procedures to obtain higher predictive performance than could be obtained from one classifier [16]. The core idea of ensemble methods is to combine several classifiers to exploit the power of each single algorithm used to obtain a more powerful classifier. Ensemble learning methods are mainly helpful if a problem can be split into subproblems so that each subproblem can be assigned to one module of the ensemble. Depending on the structure of the ensemble approach, each module can include one or more of the ML algorithms. During network attacks, because the signatures of different attacks are distinct from each other, having different sets of features as well as different ML algorithms to detect different types of attacks is preferable. A single IDS cannot address all types of input data or identify different types of attacks [17, 18]. Many researchers have shown that a classification problem can be solved with high accuracy when using ensemble models instead of single classifiers [19, 20, 21, 22].

### III. REAL CLOUD-BASED DATASET

#### A. Dataset Collection Setup

To collect real attack traffic, a testbed was built on AWS (Amazon Web Services) and was run for 10 days between the 8th and the 18th of March 2021. The system consists of three main subsystems: the Sensors, which is used as a decoy to lure the adversaries to try the system, The Collector, which gathers the data from different sensors, and the Visualizer, which parses, analyzes, searches, and extracts the collected data.

1) *The sensors subsystem*: Sensors are servers that are intentionally exposed to the public network, pretending to offer something interesting for the attacker. A lot of effort has been made to create such technology leading to what is known as a honeypot [23]. Which is a data framework asset whose esteem lies in unauthorized or unlawful utilization of that asset, which means that honeypots derive their values from the threats using them [24]. Honeypots, as a security approach, differ from firewalls and intrusion detection systems in the sense that they are implemented somewhere in the network intentionally with the hope of attracting hackers. If they are built the right way, with the right precautions, then the more they are attacked and the smarter those attacks are, the more valuable the honeypots are. A honeynet is a collection of high interaction honeypots on a tightly controlled and highly monitored network. A honeypot can be one of the three types:

- Low-interaction honeypot - This kind of honeypot gives the intruders the illusion that the system is running some services so that it has no risks and requires fewer resources, but it is easily discovered by the attacker [25].
- Medium interaction honeypot - This kind is a little more interactive as it simulates some services and enables the attacker to run commands on the system [25].
- High interaction honeypot - This kind can be a separate network of real running services for the sole purpose of deflecting the attacker from the actual services, collecting his data, and studying his behavior. It requires more resources and can be risky, but the collected data can be more valuable [26].

Besides using the honeypots as decoys to capture the attacker's data, they can also be of great value to trick and deflect the adversaries from the actual system, giving the administrators of the attacked system more time to harden the system and apply the necessary patches. In real enterprise systems, honeynets can be deployed either before or after the organization's firewall. When deployed before the firewall, they allow the most exposure to as many attacks as possible. On the other hand, they can be deployed behind the firewall for two reasons: first, to capture internal attacks originating from inside the organization by those who are trying to do things they should not be doing. This is important since internal traffic usually does not go through the firewall. Second, to give an early alert that the organization's firewall or IDS might need to be tuned after it was

successfully evaded by some non-legitimate attacker. In this experiment, the sensors subsystem is built as a honeynet of six honeypots to collect data from the attackers. Different honeypots have different purposes and run/simulate different services. In this section, a brief description of each honeypot is given.

- Dionaea: a low-interaction honeypot that captures attack payloads and malware. Dionaea listens on many different protocols, e.g., blackhole, epmap, ftp, http, memcache, mirror, mqtt, mssql, mysql, pptp, sip, smb, tftp, upnp.
- Cowrie: a medium/high-interaction honeypot that emulates SSH and Telnet services and gives the intruder the illusion of interacting with a real system and hence captures his actions against the system, e.g., commands and downloaded files. It works by running a fake filesystem with the ability to add/remove files where a full fake filesystem resembling a Debian 5.0 installation is included. It allows the addition of fake file contents so the attacker can "cat" files, such as /etc/passwd. Cowrie also gives the attacker the ability to download and upload files using wget/curl or sftp/scp and saves files downloaded for later inspection.
- Conpot: a low-interaction honeypot that is designed to work as a server-side industrial control system (ICS).
- AMUN: another low-interaction honeypot designed to capture malware that exploits server-based vulnerabilities. AMUN simulates a lot of protocols like RDP, SMP, telnet, FTP, and more to emulate many vulnerabilities e.g., Buffer Overflow, Buffer Overrun, and Stack Overflow.
- Snort: a honeypot, but it is used in this project as a sensor for the traffic. Snort is an intrusion prevention system that was developed by Cisco, who opened its source and made it available to the community.
- P0f: a tool that utilizes an array of sophisticated, purely passive traffic fingerprinting mechanisms to identify the players behind any incidental TCP/IP communications (often as little as a single normal SYN) without interfering in any way. P0f can recognize the operating system, measurement of system uptime, distance, and link type.

2) *The collector subsystem*: We used Modern Honey Network (MHN), a central server for the management and data gathering of honeypots [27]. MHN is the brain of the testbed as it facilitates the deployment of honeypots by wrapping all the necessary software for each honeypot in a script, collects data from sensors, and enables integration with the visualizer, as well as providing a RESTful API for integration with 3rd parties.

As shown in Fig. 1, MHN composed of two main components:

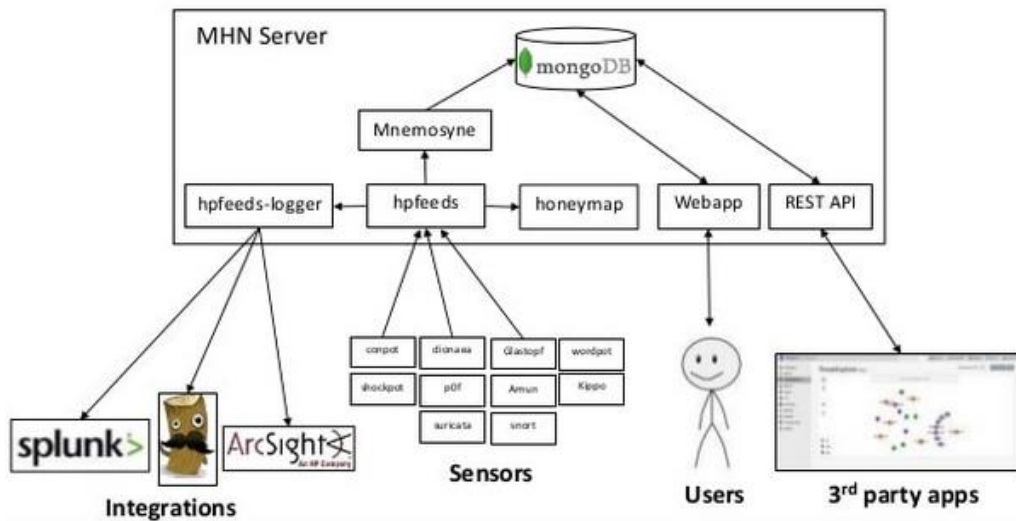


Fig. 1. MHN Server Architecture.

3) *Lightweight authenticated publish-subscribe protocol (hpfeeds)*. It has a simple wire-format so that everyone is able to subscribe to the feeds with their favorite language in very little time, so it is used as the landing point from all the honeypots, and as a data source for three other system components:

a) *Honey map*, which is a fancy map to show the geographical location of live attacks from some types of honeypots like Dionaea.

b) *Hpfeeds-logger* is a simple utility for logging hpfeeds events to files compatible with Splunk and ArcSight.

c) *Mnemosyne* provides immutable persistence for hpfeeds. It also provides normalization of data to enable sensor agnostic analysis and exposes this normalized data through a RESTful API.

4) *MongoDB is a general-purpose*, document-based, distributed database used to store all the indexed data feed from Mnemosyne. The Mongo database is used as the data source for two other system components:

a) *Web app*, which is the basic built-in visualization component of MHN unless a more complex analysis is needed by a 3rd party like Splunk.

b) *3rd party API*, which provides an API interface for 3rd party integration.

We built a testbed that consists of six sensors running different honeypots and one server running MHN and Splunk services. The honeypot servers were running on an AWS free tier t2-micro instance type, while the MHN & Splunk servers were running on a t2-medium instance during data collection, upgraded to t2-large instance type during data analysis and extraction.

5) *The Visualizer*: With the large amount of data collected by the sensors, it was better to use a third-party application to handle the data instead of the MHN built-in web app. Splunk is

used in this project, but MHN also supports integration with ArcSight software.

Splunk is a software platform to search, analyze, and visualize the machine-generated data gathered from the websites, applications, sensors, and devices that make up the IT infrastructure and business. Splunk is a great tool when it comes to the processing of a huge amount of data, as it can provide real-time processing and accept any data input format, e.g., csv, and JSON. It can also be configured to give alerts about the machine's states and predict if resource scaling is needed. To make integration with other systems easy, Splunk has the concept of apps that are an extension/addon of Splunk functionality. This gives the developers of any applications, e.g., MHN, who want to use Splunk the ability to develop their own application with a customized user interface and visualization dashboards to serve a specific need. They may then upload it to the Splunk marketplace (splunkbase) to make it available for the Splunk community. This makes it easy for the users to integrate those applications with Splunk by just importing the application extension into Splunk, and occasionally doing a few setup steps like licensing and data source configuration. For MHN, there is an app with the same name that can be downloaded from the splunkbase.

### B. The Dataset Collection Results

After the data was collected from the sensors by the MHN server and sent to Splunk for analysis and visualization, we used the Splunk query language, Splunk processing language (SPL), to extract the datasets. Table I summarizes the total amount of the collected data using the sensors subsystem, as well as the data collected per each sensor. In the section below, we present a sample of the dataset, a distribution of the data across the collection period, and a summary of the collected data per sensor.

By implementing a testbed hosted on Amazon's AWS cloud, we ran an experiment for 10 days and collected different attacks on different services. Using the data collected by different sensors, we created a real network attack dataset comprising



many interesting features that can be used to profile the attacker, e.g., source/destination IPs, source/destination port numbers (attacked service), ssh version, operating system name and version, link type, usernames and password tried by the attacker, tcp flags, ip ttl, and many more. A full list of the extracted features is shown in Table II. The dataset obtained can be used, or can be a seed, for a dataset that solves most of the common issues in the currently available datasets. It is real-world data by design, up-to-date, and can be kept up to date easily by running the testbed during specific periods. It can automate all the post-processing operations needed to get a ready dataset, thanks to the use of visualizers and query languages. The dataset represents different types of attacks and can easily represent more by deploying more honeypots.

Based on a 10-day experiment, the most attacked service was server message block (SBM), which might make sense as this service is used by the WannaCry attacks that have been spreading and active since 2017. SSH service comes second in the most attacked services as the attacker tries to exploit the lack of awareness of some users that use the default or weak credentials. The common username, “admin,” was the most tried username and “password” came second as the most tried passwords, while the less expected, “nproc” (a bash command to get the total number of cores/threads on the machine) was the most tried password. The most used operating system by the attacker was Linux version 3 or later, while Windows came next, which makes sense as a lot of the hacking tools used are Linux-based, e.g., Kali. Although most of the attacks originated from the United States, it might or might not accurately reflect the actual attacker’s location since a serious attacker might be using compromised machines to mount his attacks. These could be located anywhere, or be using any cloud-hosted machines, which the US has most. The data showed that the top attacking single IP was in Panama and generated around 34,000 attacks during the 10 days. The most used command is “uname,” which is used to get the operating system type, kernel version, and other information that is necessary to determine the suitable attacking scripts and tools. The second and the third most used commands are “echo” which is used to show whatever argument is used after it and “which ls” to get the full path of the “ls” command. The two commands might not be meant for actual use but just to check if this is a real system or a trap. This is good to know as it can guide the honeypot developers toward which commands, they need to simulate for a more deceptive honeypot.

TABLE I. TOTAL COLLECTED DATA

Sensor	Total Collected	Distinct SRC	Distinct SRC DEST_Port	Distinct SRC DEST_Port SRC_Port
Dionaea	177,000	10,000	72,000	158,000
Pof	369,000	24,000	108,000	212,000
AMUN	245,000	9,000	10,000	228,000
Cowrie	58,000	1,243	1,243	45,000
Snort	108,000	6,200	53,000	67,000
Conpot	3,780	444	444	544
Total	960,780	50,887	244,687	755,544

TABLE II. THE FULL LIST OF EXTRACTED FEATURES

#	Feature Name	Description
1	_time	time of traffic capturing
2	app	honeypot captured the traffic
3	dest	dest ip
4	dest_port	dest port
5	dionaea_action	either Dionaea honeypot accept or reject the connection
6	direction	the direction of the captured traffic either in or out
7	eth_dst	the dest mac address
8	eth_src	the source mac address
9	host	Splunk server ip or hostname
10	ids_type	the type of the used ids
11	ip_id	the packet id
12	ip_len	packet length
13	ip_tos	packet type of service
14	ip_ttl	packet time to live
15	linecount	the number of lines of the captured traffic
16	p0f_app	protocol used by P0f for fingerprinting
17	p0f_link	the connection type at the attacker side like modem or dsl
18	p0f_os	the operating system of the machine generating the attack
19	p0f_uptime	how long since the attacking machine is up
20	protocol	tcp or udp
21	sensor	id assigned by MHN per honeypot
22	severity	severity rank of the attack
23	signature	the signature of the attack as matched by snort
24	snort_classification	a number given by snort to classify the traffic
25	snort_header	the rule header
26	snort_priority	assigns a severity level to rules
27	source	input data source (needed by Splunk)
28	sourcetype	input data type (needed by Splunk)
29	splunk_server	Splunk ip or hostname
30	src	attack src ip
31	src_port	attack source port
32	ssh_password	password used by the attacker trying to get ssh access
33	ssh_username	username used by the attacker trying to get ssh access
34	ssh_version	attacker ssh client version
35	tcp_flags	indicate a particular connection state or provide additional information
36	tcp_len	packet length
37	timeendpos	at which byte into the event the timestamp ends
38	timestartpos	at which byte the timestamp starts
39	transport	transport protocol type tcp or udp
40	type	honeypot event type
41	udp_len	packet length
42	vendor_product	name of the honeypot that captures the traffic
43	_raw	raw (not parsed) event

Analysis of the collected data showed some interesting findings for each sensor. Dionaea not only listens on the opened ports but also allows the attacker to download and upload files. Fig. 2 shows a list of the top downloaded binaries expressed as their MD5. It is worth noticing that the same files came from different sources. Splunk's MHN application also adds a fast method to scan those files against different antiviruses via Total Hash and Virus Total websites. By clicking on the link, a web page will open automatically, search for the file, and show the scanning results, as shown below in Fig. 3. Fig. 4 to 8 show the most often attacked ports in different sensors. Fig. 9 shows the top link types, and Fig. 10 shows the operating systems. Fig. 11 shows the top URLs that were used by the attackers to download scripts, and binaries used to mount their attacks. Fig. 12 shows the top SSH versions, Fig. 13 shows the most used user/password pairs, and Fig. 14 shows the most often used attack commands. Fig. 15 shows the top attack types captured by Conpot.

#### IV. STACKING ENSEMBLE IDS

This section presents an ensemble learning model using a meta-classification method enabled by stacked generalization. A newly generated dataset that was captured from real network traffic was used for experimentation. Observed results indicate that the proposed stacking ensemble can generate superior predictions having 95% accuracy.

##### A. Methodology

As illustrated in Fig. 16, the stacking model comprises base and meta-classifiers—namely, Neural Networks (NN), k nearest neighbor (KNN), Decision Tree (DT), and Support Vector Machine (SVM), respectively.

Authors in [28] illustrated that the integration of a set of single algorithms leads to optimum predictions. Stacking or stacked generalization is a concept proposed by Wolpert [29]. Several ML algorithms define their subjective biases on a learning set, ultimately filtering out biases. The implementation of a stacked model involves two kinds of sub-models, base (level 0 classifiers) and metamodels (level 1 or meta). The main logic of a stacking model lies in using the meta-classifier to predict the samples by studying the level 0 classifiers. Yan and Han [30] illustrated the great advantage of using the stacking models. They have stated that stacking can enhance prediction accuracy while working with unbalanced datasets. A study [31] was conducted to emphasize the application of AI-based classifiers. The researchers in that study explained that ensembles were able to adapt to the robust behaviors of malicious and normal traffic effectively. Algorithm 1 shows the entire classification process implemented in the classification framework involving multiple classifiers.

##### B. Data Pre-processing and Feature Selection

Pre-processing was utilized to handle different data found in the dataset. To eliminate noise, and fix inconsistencies found in the data, a statistical transformation tool is needed. In our proposed work, missing data and outliers were compensated for by making the distribution normal. However, lost values rely on singular features. While some features can be assigned zero as a missing value, others are assigned zero as an actual value where binary data are considered. To maintain such predicaments, consideration of relevant features that guarantee ideal expectations is essential. Thus, an integration of hashing and information gain (IG) was applied to extract the maximum desirable features. Feature scaling was also utilized to assure that those features possessing a greater numeric range did not dominate the ones in smaller numeric ranges. The dataset has many features but not all appear to be important. Consequently, only 11 features were chosen from the dataset. The fundamental features were assigned weights to prioritize them, and only the best features were extracted. The dimensionality of the features was reduced using a hashing approach. The chosen features are direction, eth\_src, host, protocol, src, src\_port, ssh\_version, tcp\_flags, tcp\_len, type, and udp\_len.

##### C. Classification

The methodology to actualize the classification system included the application of different classifiers to resolve the basic complexities of information found in both packet-based and flow-based datasets.

Fundamentally, KNN count on a distance function that calculates the similarity or variance between two network occurrences found in the datasets under consideration.

The Euclidean distance  $d(x, y)$  can be calculated by via the following equation:

$$d(x, y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2} \quad (1)$$

where  $x_i$  is the  $i$ th feature of the instance  $x$ , while  $y_i$  is the  $i$ th feature of the instance  $y$ , and “ $n$ ” is the whole number of features found in the dataset. Let  $C = C_1, C_2, C_3, \dots, C_p$ . There are “ $p$ ” labels in the dataset. Let “ $x$ ” be the new sample to be predicted. The objective of KNN classifier is to determine “ $k$ ” vectors that are close to  $x$ . If most of the vectors belong to class  $C_m$ , then  $x$  will be assigned the class label  $C_m$ .

The radial basis function (RBF) is a preferred kernel function for many classification problems in ML. The following equation defines the RBF:

$$k(x, y) = \exp \left( -\frac{\|x - y\|^2}{2\sigma^2} \right), \quad (2)$$

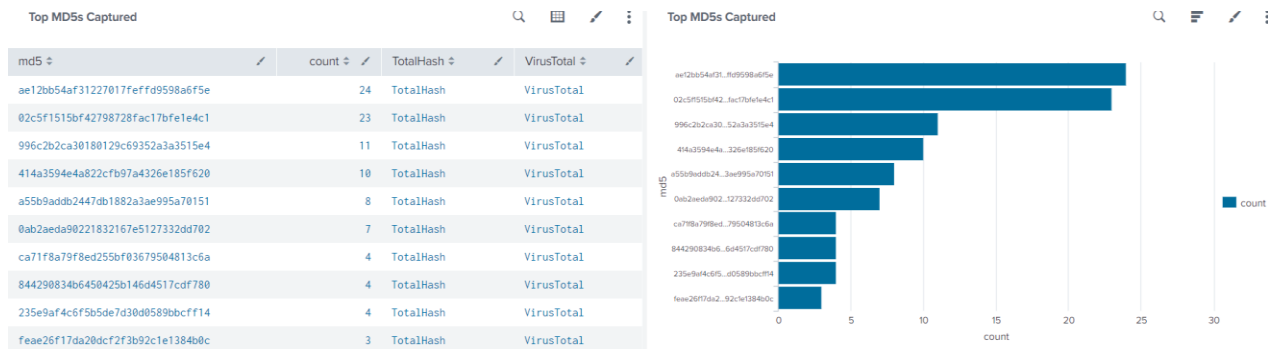


Fig. 2. Dionaea Top Captured MD5Binaries.

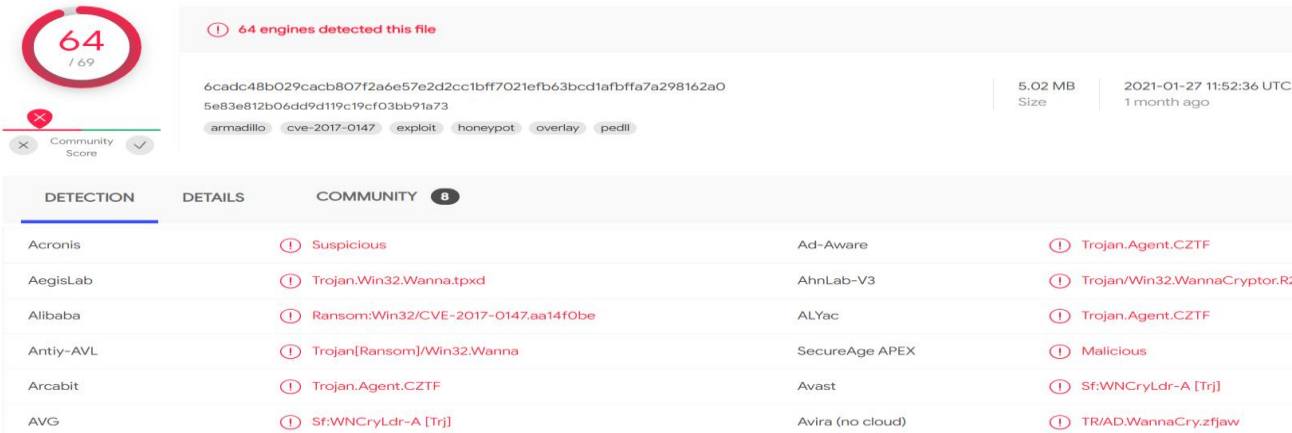


Fig. 3. Scanning Results for a Malware File.

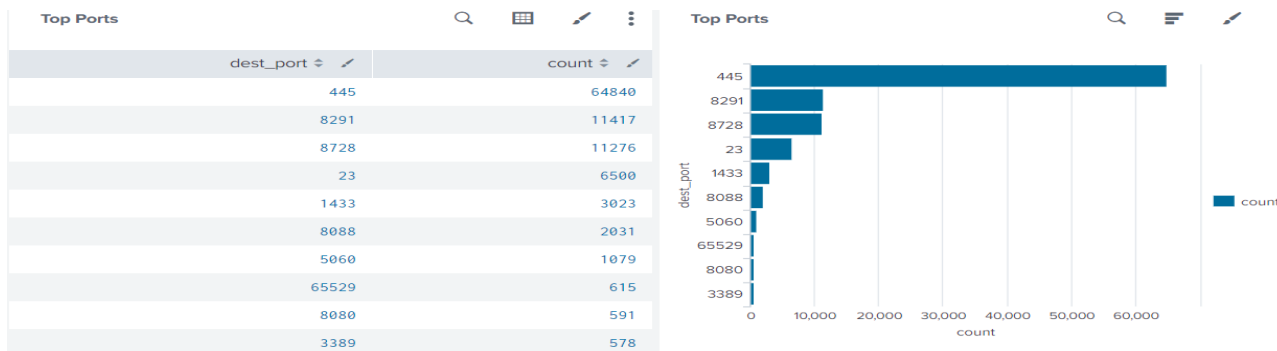


Fig. 4. Dionaea Top Attacked Ports.

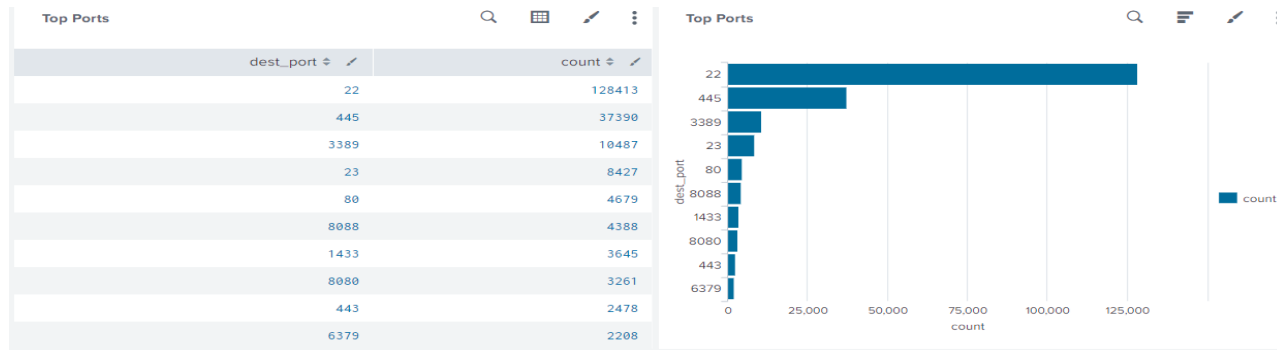


Fig. 5. POF Top Attacked Ports.

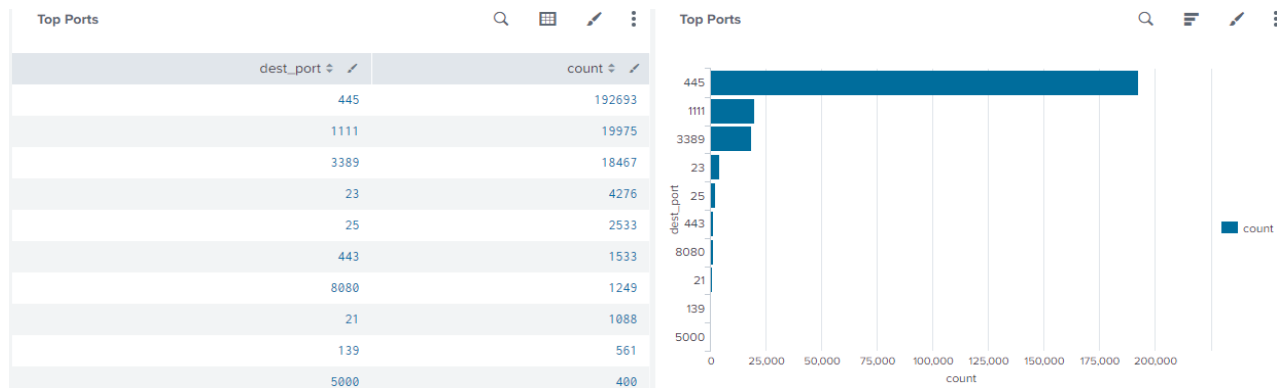


Fig. 6. AMUN Top Attacked Ports.

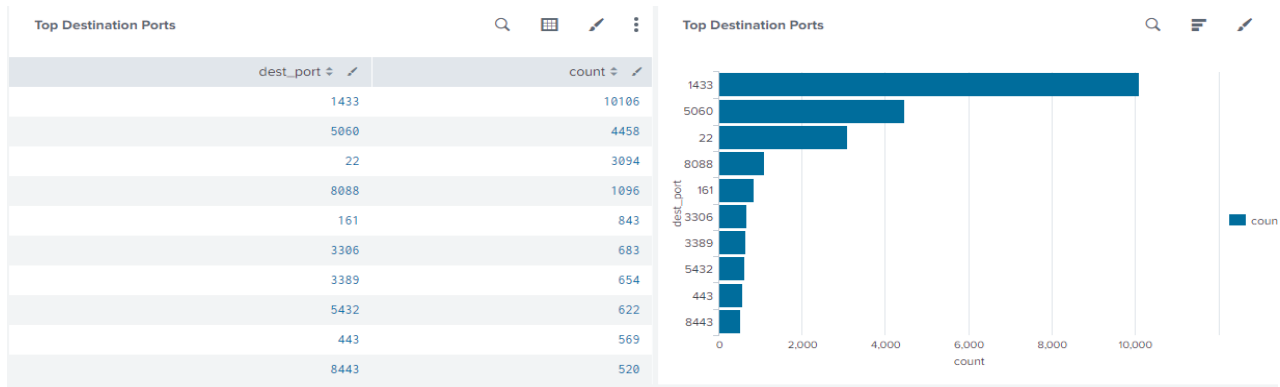


Fig. 7. Snort Top Attacked Ports.

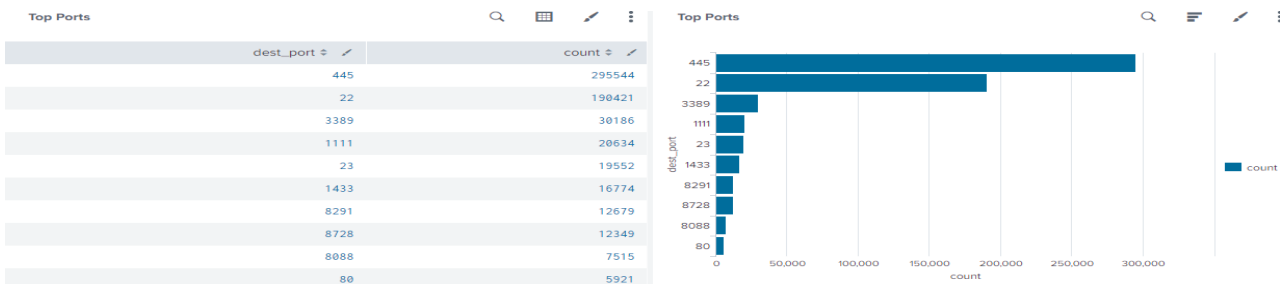


Fig. 8. Most Attacked Ports.

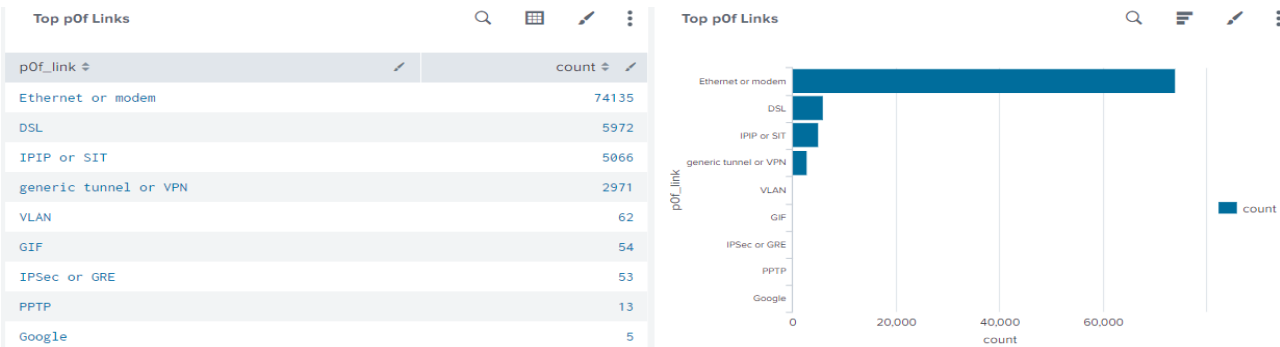


Fig. 9. PoF Top Link Types.

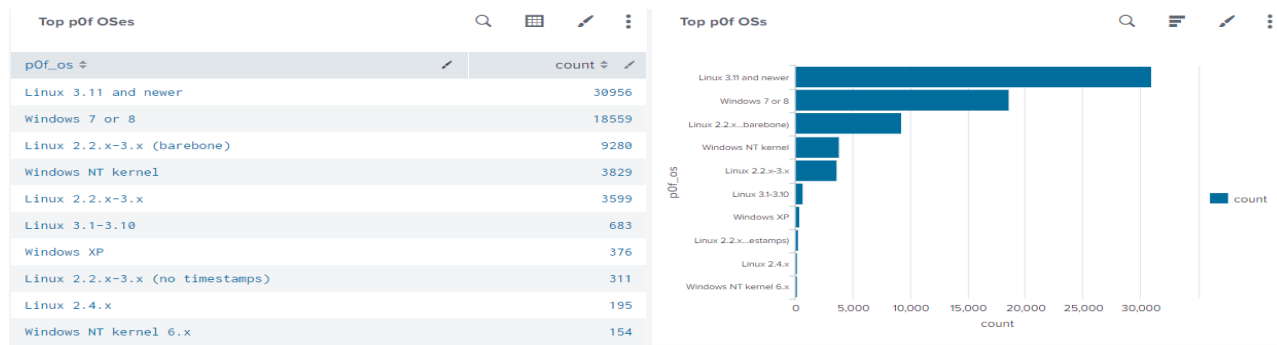


Fig. 10. POf Top Operating Systems.

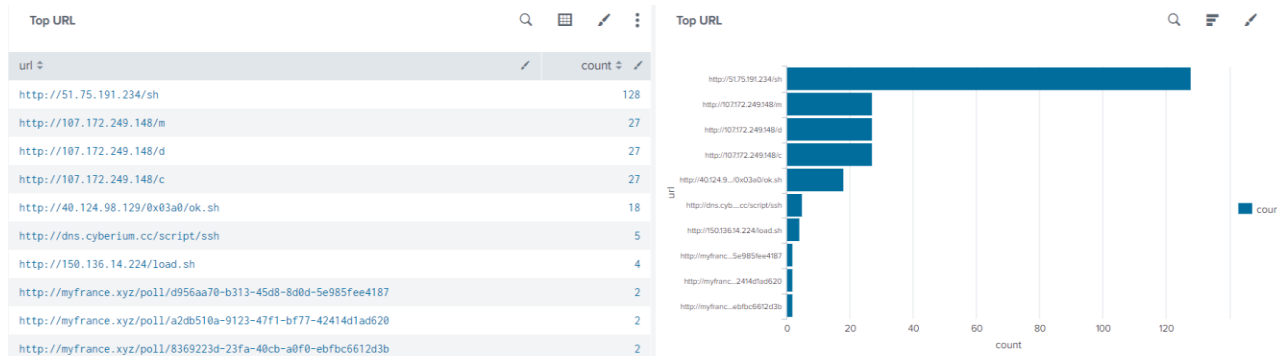


Fig. 11. Cowrie Top URLs.

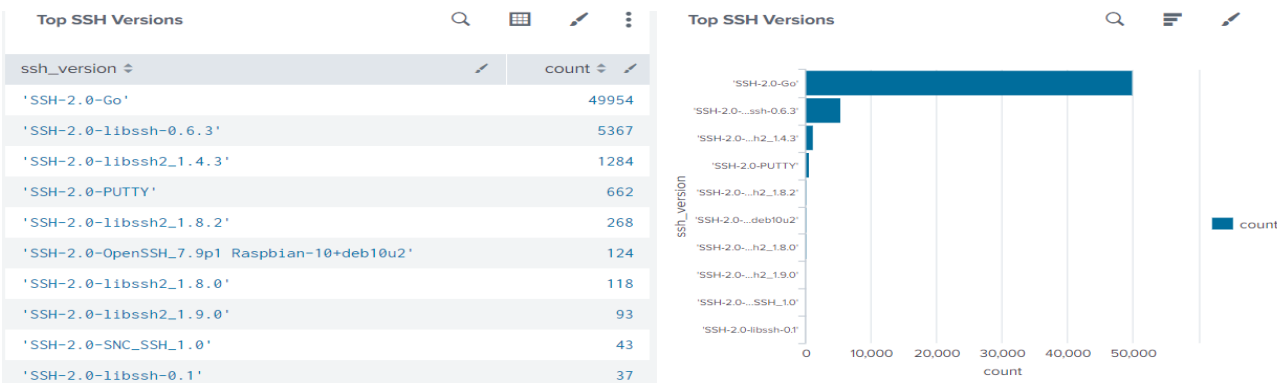


Fig. 12. Cowrie Top SSH Versions.

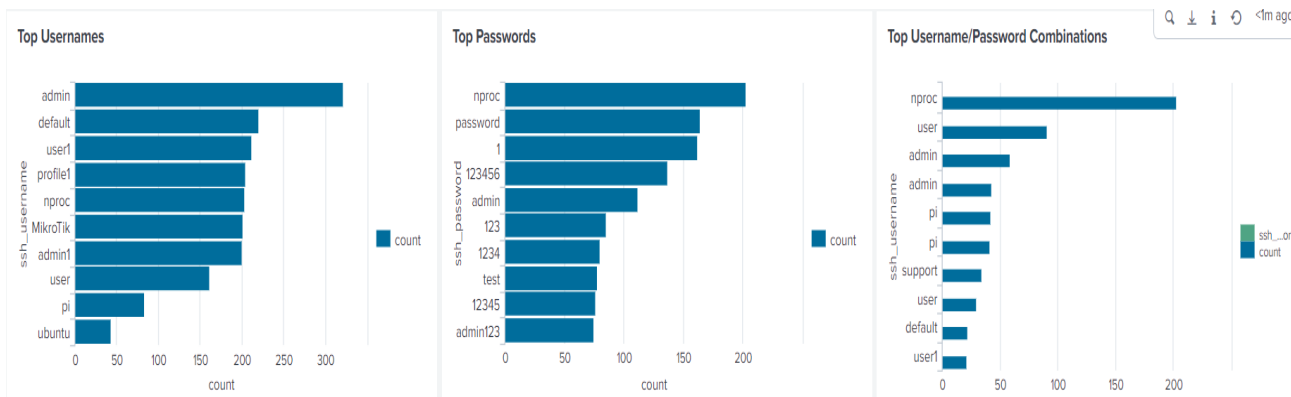


Fig. 13. Cowrie Top Users/Passwords.

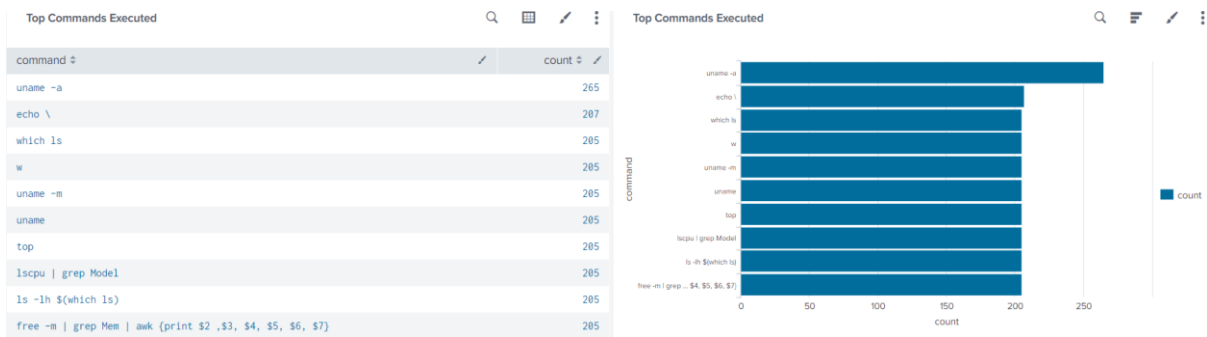


Fig. 14. Cowrie Top Attack Commands.

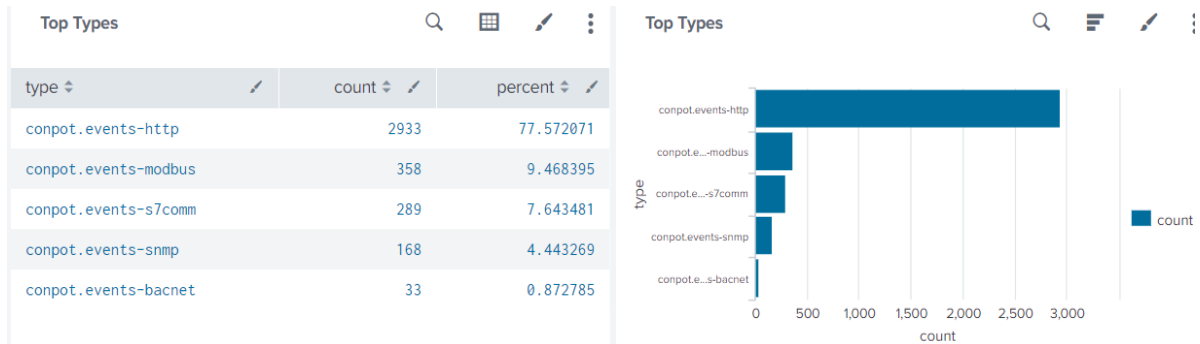


Fig. 15. Conpot Captured Top Attack Types.

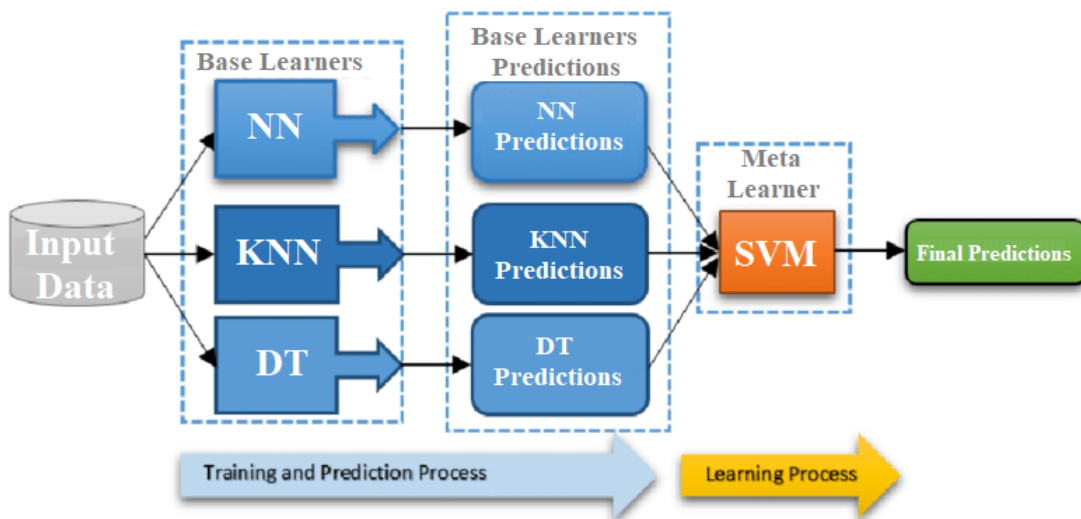


Fig. 16. Stacking Ensemble Model.



Algorithm 1: Stacking Ensemble Strategy.

**Input:** Train data  $T = \{X_i, Y_i\}_{i=1}^m$  ( $X_i \in R^n, Y_i \in Y$ )  
**Output:** Predictions from the ensemble E  
*Step 1.* Impose cross validation in order to prepare a training set for meta-classifier  
*Step 2.* Randomly split  $T$  into “ $m$ ” equal size subsets, i.e.,  $T = \{T_1, T_2, T_3 \dots T_m\}$   
*Step 3.* for  $m \leftarrow 1$  to  $M$   
    Learn base classifiers namely NN, KNN, and DT  
    for  $n \leftarrow 1$  to  $N$   
        Learn a classifier  $P_{mn}$  from  $T$  or  $T_m$   
    End for  
*Step 4.* Formulate a training set for metaclassifier (SVM)  
    for each  $X_i \in T_m$   
        Extract a new instance  $(x'_i, y_i)$ , where  $x'_i = \{P_{m1}(X_i), P_{m2}(X_i), P_{m3}(X_i), \dots, P_{mN}(X_i)\}$   
    End for  
End for  
*Step 5.* Return  $y_{\bar{1}} = \{y_1, y_2, y_3, \dots, y_n\}$  from ensemble

Algorithm 1 shows the whole classification process implemented in the classification framework involving multiple classifiers.

D. Results and Discussion

The quality of any IDS can be measured by four performance metrics: true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN).

To accurately value the performance of the proposed approach and assure the results acquired from the stacked ensemble model, both binary and multiclass classification results are given in this section. Table III depicts the results acquired upon classifying the network instances of the dataset into either attack or normal. Moreover, to test the predictions and to assure that the models do not overfit, mean training accuracy (MTA), mean training precision (MTP), and mean training recall (MTR) values are also mentioned in Table IV.

The NetFlow traces found in the dataset contain genuine background network traffic for a substantial duration of ten days. As per the confusion matrix presented in Table V, all seven attack types found in the dataset were distinguished perfectly aptly by the stacking classifier.

The proposed ensemble model could identify the occurrence of SSH scan attack in the foremost effective way. In order to show reliable results, performance metrics like precision and recall were also considered in addition to accuracy. Recall is the ability of the intrusion detection model to correctly locate the positive instances, where precision is the model's capability to locate the percentage of positive instances that were identified correctly.

Table VI shows that the false alarm rate is extremely least regarding all attack classes, which is a signal that the general effectiveness of the proposed ensemble model is very good. The ROC curve also is shown in Fig. 17.

TABLE III. BINARY CLASSIFICATION RESULTS

Accuracy	Precision	Recall	F1 score	AUC	FAR (%)
0.94	0.96	0.93	0.95	0.99	5.2

TABLE IV. TRAINING RESULTS OBTAINED BY 10-FOLD CROSS VALIDATION

Fold Number	Training Accuracy	Training Recall	Training Precision
1	0.9289	0.9142	0.9488
2	0.9299	0.9129	0.9520
3	0.9239	0.9101	0.9469
4	0.9278	0.9102	0.9500
5	0.9301	0.9109	0.9531
6	0.9319	0.9129	0.9480
7	0.9430	0.9040	0.9481
8	0.9258	0.9089	0.9519
9	0.9290	0.9140	0.9479
10	0.9260	0.9110	0.9509
	MTA: 0.9285	MTR:0.9115	MTP: 0.9497

TABLE V. CONFUSION MATRIX OF ALL THE 7 ATTACK TYPES

		0	1
SSHscan	0	0943204	05809
	1	07824	091829
UDPscan	0	0891295	016466
	1	018477	0121338
Spam	0	0932187	09632
	1	013268	093589
DOS	0	0936949	013311
	1	09348	089968
Scan	0	0927854	011710
	1	010253	099759
Blacklist	0	0940476	09738
	1	08962	089420
DDOS	0	0927785	014583
	1	060303	046915

TABLE VI. CLASS-WISE PERFORMANCE

Metric	Blacklist	Spam	Scan	SSH scan	UDPscan	DO S	DD OS	Overall
Recall	0.9918	0.98597	0.98907	0.99056	0.9797128	0.99012	0.93897	0.9809
Precision	0.9940	0.98988	0.98859	0.98976	0.9818808	0.98703	0.98557	0.9881
FAR	0.0054	0.0062	0.0102	0.0093	0.0147	0.0117	0.0130	0.0101
Accuracy	0.9871	0.97826	0.98001	0.98218	0.9666758	0.97934	0.92954	97.19%

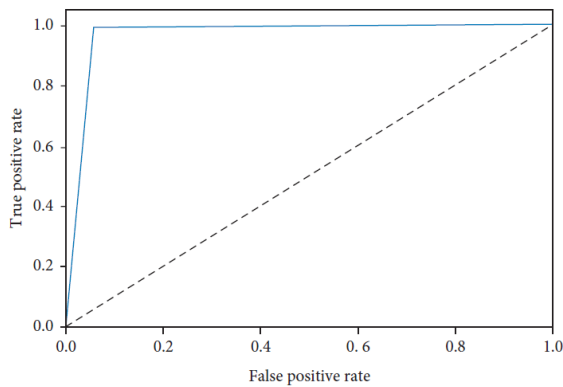


Fig. 17. ROC Curve.

## V. CONCLUSION AND FUTURE WORK

This paper has presented an ensemble methodology based on a newly generated dataset that was extracted from real network traffic. The extensive dataset that has been created provides valuable benefits for training ML models to detect current attack types efficiently and accurately. This is because it overcomes most deficiencies of the present available datasets and covers most of the essential standards with common updated attacks. Moreover, the produced dataset is fully labeled and includes different network traffic features that are extracted and calculated for all normal and intrusion flows.

To utilize the created dataset, we presented an adaptive stacking ensemble learning model that integrates the advantages of different ML algorithms for diverse kinds of attacks and achieves optimal results through ensemble learning. This combines the predictions of several base estimators (i.e., NN, KNN, DT, and SVM) to accelerate the processing speed and improve scalability with a larger amount of network traffic data. The experimental results have shown that the ensemble model was able to enhance the classification accuracy, increase the true positive rate, and decrease the false positive rate. The real dataset provided can help cybersecurity researchers and firms to better understand the recent networking environment traffic, and traits of recent attacks, in order to better detect and prevent them. It can also help law enforcement and digital forensics teams in investigating cyberattacks. The proposed ensemble model can also be utilized with the provided dataset as a training dataset to detect and classify potential network attacks. This can help service providers, like cloud service providers, to monitor and improve their infrastructure.

This work can be expanded in the future to cover more and/or new attacks by collecting more networking traffic in different environments such as the Internet of Things networks, fog, etc. In addition, we can investigate the effectiveness of the ensemble model against such new networking traffic and suggest different features and tuning for every type of environment. More experimental analysis and a complete comparison with literature would be considered as well.

## VI. DATA AVAILABILITY

The dataset generated in this work is publicly available and can be accessed from this link. [https://www.researchgate.net/publication/356809493\\_Towards\\_A\\_Holistic\\_Efficient\\_Stacking\\_Ensemble\\_Intrusion\\_Detection\\_System\\_Using\\_Real\\_Cloud-based\\_Dataset](https://www.researchgate.net/publication/356809493_Towards_A_Holistic_Efficient_Stacking_Ensemble_Intrusion_Detection_System_Using_Real_Cloud-based_Dataset).

## REFERENCES

- [1] Libert, B., M. Beck, and J. Wind, The network imperative: How to survive and grow in the age of digital business models. 2016: Harvard Business Review Press.
- [2] Neumann, P.G., Computer-related risks. 1994: Addison-Wesley Professional.
- [3] Demestichas, K., N. Peppes, and T.J.S. Alexakis, Survey on security threats in agricultural IoT and smart farming. 2020. 20(22): p. 6458.
- [4] Cheminod, M., L. Durante, and A.J.I.t.o.i.i. Valenzano, Review of security issues in industrial networks. 2012. 9(1): p. 277-293.
- [5] Gorelik, E., Cloud computing models. 2013, Massachusetts Institute of Technology.
- [6] Haq, N.F., et al., Application of machine learning approaches in intrusion detection system: a survey. 2015. 4(3): p. 9-18.
- [7] Moustafa, N. and J. Slay. The significant features of the UNSW-NB15 and the KDD99 data sets for network intrusion detection systems. in 2015 4th international workshop on building analysis datasets and gathering experience returns for security (BADGERS). 2015. IEEE.
- [8] Shiravi, A., et al., Toward developing a systematic approach to generate benchmark datasets for intrusion detection. 2012. 31(3): p. 357-374.
- [9] Li, Y.-F., et al., A systematic comparison of metamodeling techniques for simulation optimization in decision support systems. 2010. 10(4): p. 1257-1273.
- [10] Khraisat, A., et al., Survey of intrusion detection systems: techniques, datasets and challenges. 2019. 2(1): p. 1-22.
- [11] Xie, H., K. Lv, and C. Hu. An effective method to generate simulated attack data based on generative adversarial nets. in 2018 17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/12th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE). 2018. IEEE.
- [12] Grajeda, C., F. Breiteringer, and I.J.D.I. Baggili, Availability of datasets for digital forensics—and what is missing. 2017. 22: p. S94-S105.
- [13] Devi, M.G. and M.J. Nene. Scarce Attack Datasets and Experimental Dataset Generation. in 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA). 2018. IEEE.
- [14] Kholidy, H.A. and F. Baiardi. Cidd: A cloud intrusion detection dataset for cloud computing and masquerade attacks. in 2012 Ninth International Conference on Information Technology-New Generations. 2012. IEEE.
- [15] Nazarov, A., A. Sychev, and I. Voronkov. The Role of Datasets when Building Next Generation Intrusion Detection Systems. in 2019 Wave Electronics and its Application in Information and Telecommunication Systems (WECONF). 2019. IEEE.
- [16] Dasgupta, D., et al., Machine learning in cybersecurity: a comprehensive survey. 2020: p. 1548512920951275.
- [17] Shalev-Shwartz, S. and S. Ben-David, Understanding machine learning: From theory to algorithms. 2014: Cambridge university press.
- [18] Jordan, M.I. and T.M.J.S. Mitchell, Machine learning: Trends, perspectives, and prospects. 2015. 349(6245): p. 255-260.

- [19] Fayyad, U.M. and K.B.J.M.I. Irani, On the handling of continuous-valued attributes in decision tree generation. 1992. 8(1): p. 87-102.
- [20] Diplaris, S., et al. Protein classification with multiple algorithms. in Panhellenic Conference on Informatics. 2005. Springer.
- [21] Oza, N.C. and K.J.I.f. Tumer, Classifier ensembles: Select real-world applications. 2008. 9(1): p. 4-20.
- [22] Chand, N., et al. A comparative analysis of SVM and its stacking with other classification algorithm for intrusion detection. in 2016 International Conference on Advances in Computing, Communication, & Automation (ICACCA)(Spring). 2016. IEEE.
- [23] Perkins, R.C. and C.J. Howell, Honeypots for Cybercrime Research, in Researching Cybercrimes. 2021, Springer. p. 233-261.
- [24] Spitzner, L. Honeypots: Catching the insider threat. in 19th Annual Computer Security Applications Conference, 2003. Proceedings. 2003. IEEE.
- [25] Almotairi, S., et al. A technique for detecting new attacks in low-interaction honeypot traffic. in 2009 Fourth International Conference on Internet Monitoring and Protection. 2009. IEEE.
- [26] Nicomette, V., et al., Set-up and deployment of a high-interaction honeypot: experiment and lessons learned. 2011. 7(2): p. 143-157.
- [27] Wafi, H., et al. Implementation of a modern security systems honeypot honey network on wireless networks. in 2017 International Young Engineers Forum (YEF-ECE). 2017. IEEE.
- [28] Van der Laan, M.J., et al., Super learner. 2007. 6(1).
- [29] Wolpert, D.H.J.N.n., Stacked generalization. 1992. 5(2): p. 241-259.
- [30] Yan, J. and S.J.M.P.i.E. Han, Classifying imbalanced data sets by a novel re-sample and cost-sensitive stacked generalization method. 2018. 2018.
- [31] Kumar, G., K.J.A.C.I. Kumar, and S. Computing, The use of artificial-intelligence-based ensembles for intrusion detection: a review. 2012. 2012.

# Authorship Attribution on Kannada Text using Bi-Directional LSTM Technique

Chandrika C P, Jagadish S Kallimani  
Ramaiah Institute of Technology, Bangalore-54, India  
Affiliated to Visvesvaraya Technological University  
Belagavi, Karnataka, India

**Abstract**—Author attribution is the field of deducing the author of an unknown textual source based on certain characteristics inherently present in the author's style of writing. Author attribution has a ton of useful applications which help automate manual tasks. The proposed model is designed to predict the authorship of the Kannada text using a sequential neural network with Bi-Directional Long Short Term Memory layers, Dense layers, Activation function and Dropout layers. Based on the nature of the data, we have used stochastic gradient descent as an optimizer that improves the learning of the proposed model. The model extracts Part of the speech tags as one of the semantic features using the N-gram technique. A Conditional random fields model is developed to assign Part of the speech tags for the Kannada text tokens, which is the base for the proposed model. The parts of the speech model achieve an overall 90% and 91% F1 score and accuracy respectively. There is no state-of-art model to compare the performance of our model with other models developed for the Kannada language. The proposed model is evaluated using the One Versus Five (1 vs 5) method and overall accuracy of 77.8% is achieved.

**Keywords**—Authorship attribution; Bi-Directional Long Short Term Memory; machine learning algorithms; parts of speech; stylometry features

## I. INTRODUCTION

Authorship Attribution (AA) finds the hidden patterns in an author's writing to identify the author of an unknown text. Not much work has been done for the same, especially for the texts in the Kannada language, which is a popular Indian regional language. The authorship attribution system works primarily to predict the probability of mapping the article to its author. Authorship attribution is a field with significant applications and a long history to present a solution to the same. Recent works in this domain for foreign languages have proven to be a powerful automated tool but in Indian regional languages, the absence of a state-of-the-art method leaves scope for improvement and research. Advancements in Machine learning techniques, Natural Language Processing (NLP) and Artificial Intelligence (AI) have helped in developing a model for author attribution by learning the distinct features in the author's writing style.

Kannada, the state language of Karnataka, belongs to India, is rich in literature and culture and the Kannada-speaking people are spread around the globe. Text processing is a challenging one. Deep learning algorithms [1] like Bidirectional Encoder Representations from Transformers, a transformer based model and a hybrid model [2] composed of

Recurrent Neural Network (RNN), Long Short Term Memory (LSTM), and Gated Recurrent Unit (GRU) proved effective in text classification.

Text Classification necessitates a significant amount of time spent analyzing the contents [3]. Several parameters like large vocabulary, semantic ambiguity, and words having meaningful relationships are used to classify the text.

Now-a-days text processing in Kannada is rapidly growing, AA can be found useful in predicting the ownership of a Kannada disputed text like threat letters, suicidal notes, literature work and so on. As per our knowledge, the proposed work is a novel approach. Till now no significant work has been carried out in the Kannada language on the authorship attribution of digital text. One can find a few works which emphasize handwriting analysis [4] so there is a lot of scope in this field, especially in the local languages. There are two main approaches to authorship attribution: Profile and Instance based approaches. The former is mainly suitable for short article samples and the latter is employed for lengthy articles. The proposed model uses a profile-based approach, in which the features are extracted from short samples to create an author's profile and then trained and tested with deep learning networks. To test the owner of a Kannada handwritten document, the handwriting styles [5-6] like cursive line, font size, the thickness of line, formation of characters, spaces between characters and words, and so on are considered but when the text is digital, then different parameters have to be used for the comparisons, these parameters are referred as Stylometric features. Stylometry features are those special features used to extract a person's writing style like lexical features, which include a total number of words/sentences/ special symbols/ usage of nouns and vocabulary richness, etc. Semantic features like POS tags and content-based features etc.

In our previous works [7-8], two AA models were developed based on lexical and syntactic features using classification algorithms and the N-grams technique respectively and observed that these models predict the probability of authorship pretty well. In the proposed work, semantic features are extracted using POS tags. Deep learning techniques are popular for many NLP applications. From the survey, it is found that Deep learning networks combined with N-gram is an efficient technique for many text processing applications and it improves the performance of a model to a great extent. Bidirectional Long Short Term Memory (Bi-LSTM) is the process of constructing a neural network that can

store a sequence of information in both directions forward (future to past) and backward (past to future). Inputs run in two directions in a bidirectional LSTM, which distinguishes it from a conventional LSTM.

The process for training and testing the proposed model using BI-LSTM is shown in the diagram below. The primary objectives for developing this model are:

- To extract semantic features of an author.
- To develop a Kannada POS tagger.
- To develop the Kannada AA model using deep learning techniques.
- Performance comparison of the proposed model with other languages models.

Fig. 1 shows the process of AA, the input to this model is the cleaned labeled Kannada dataset, quality features like POS and N-grams are extracted from this, which are later trained with machine learning models like Bidirectional LSTM and during the testing phase the anonymous text is questioned and the model predicts the most suitable author for the text.

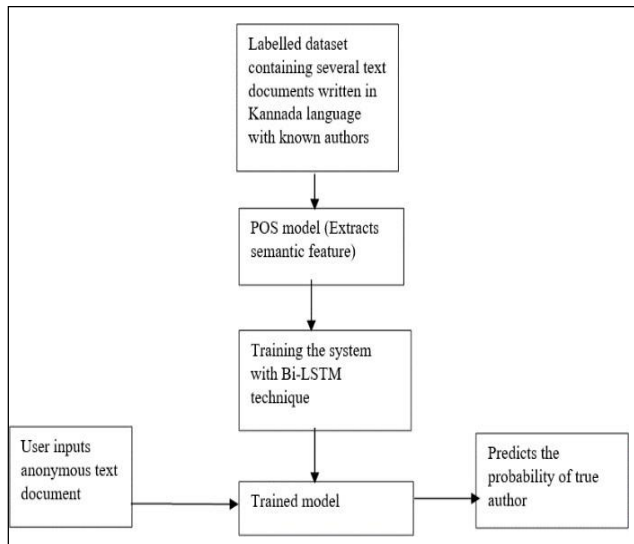


Fig. 1. Authorship Attribution Process.

### Contributions

- The proposed work focuses on predicting the authorship of an anonymous Kannada text.
- The AA model accuracy mainly depends on the quality features, apart from extracting lexical features, semantic features are also extracted using the POS model, for this a POS tagger for Kannada tokens using the CRF model is developed.
- The work demonstrates the implementation of deep learning techniques like Bidirectional LSTM and using POS and grams approaches to perform AA task.
- The proposed work considers 50 authors of 500 documents and the overall accuracy of 77.8% is achieved.

## II. LITERATURE SURVEY

A survey will help us to analyze different techniques and methodologies explored by different researchers for Authorship attribution. This section will describe the same. A detailed survey on AA is done in [9], the authors described different dimensions of Authorship analysis including authorship prediction, verification, the importance of Stylometry features, ML algorithms and Deep learning techniques on AA. This work serves as a prerequisite for a researcher to start his work in the AA domain. Authors in [10] have explored an Instance based AA using deep learning based Artificial Neural Network on the Arabic Language. ANN (the proposed solution) produced an accuracy of 75.46% compared to 68.85%, 69.78%, 69.64%, and 69.78% attained by SVM, RF, DT and BNB respectively. Authors in [11] have explored a Cross-domain AA using Character sequences, Word uni-grams, and POS-tags features. Both the first and the second model extracts char 6-gram and 3-8-grams respectively. The third model was composed of content-based features based on POS tags. The results on the evaluation corpus are significantly lower and the three models seem to be overfitting. A different technique called Life-Like Network Automata for AA tasks is explored in [12]. This research represents network modeling texts as network automata (LLNA) with dynamics based on Life-Like rules. The LLNA method searches the whole rule space for an optimal solution to one problem. The best results were obtained with a partial lemmatization process, suggesting that this procedure is more adequate than just lemmatizing all words when text networks are used as the underlying model for this task.

Instance based and Profile-based approaches based on ensemble strategy to maximize the outcome by combining the probabilities of different feature sets by using SVM are discussed in [13]. AA task experiments on four different languages, researchers have also employed SVM with linear kernel and RBF kernel, K-nearest neighbors with K=3, and Random Forest and obtained an F1 score of 68%. Deep learning techniques proved to be very effective for AA tasks. Extracting the lexical features of an author and calculating projection for each file to predict the authorship [14] was found to be interesting. The result of the average projection shows similarities between the main author file and the summary file of each author. To recognize the true author of anonymous text written in the Russian language using deep learning networks is discussed in [15]. Authors have Extracted 33 to 5000 features and then trained and tested them using SVM and other NN features like optimizing algorithms, dropouts, loss function and various activation functions. SVM performed well with 96% average accuracy compared to a Deep neural network with 93% accuracy.

A convolution neural network based authorship model for the Bengali language is demonstrated in [16], authors have considered 6 author's 350 samples, and character level pre-trained embedding called fastText gives maximum accuracy of 98%, this work proves that pre-trained embedding outperforms compared to the non-pre-trained embedding of the text. The pre-trained models like BERT, Embeddings from Language Models(ELMo), Universal Language Model Fine-tuning and generation Generative Pre-trained Transformer -2

based authorship prediction on cross domain has been demonstrated in [17], a multi-headed classifier and DEMUX layer is created to handle different classifiers, BERT and ELMo outperform with more than 90% accuracy compared to other language models. Stylometry features play a vital role in the AA task, authors in [18] have explored a new technique of generating human-like sentences using a neural network and then various linguistic features are extracted to predict the authorship. the proposed model with an accuracy of 97.2% can predict the true author successfully.

AA for a very lengthy corpus is a tedious job, using a reduction model [19], the size of the candidate authors can be reduced. Doc2Vec reduction was found to be efficient compared to other models used for reduction. It is observed that Reduction in the candidate authors set and the corpus didn't significantly affect the performance of the AA model. Reduction of authors set with a minimum of 10% and a maximum of 90%, the model achieved 99% and 50% accuracy respectively. Lexical feature extraction is easy compared to semantic or content based features, but semantic features are more realistic. Researchers in [20] developed a content based model for AA by considering authors from different domains and also datasets with different genres. The proposed model learns the sentences from POS tags and the system is trained and tested with RNN, the model was able to achieve maximum accuracy of 78% accuracy on the PAN dataset. AA on Persian historical and literary works is explored in [21]. The authors have used a modified four parts of deep convolutional neural networks architecture and attention mechanism. The model outperforms other approaches with 72.59% accuracy.

Authors [22] have tried to predict the owner of the e-mail which has some dispute contents, a user with fake mail ids can write unacceptable contents that may damage the reputation of a person/ company. Prediction is mainly based on reasonable hypotheses; authors have strived to develop a mathematical model to successfully address this problem by combining the Analytic Hierarchy Process with SVM. Experimental findings demonstrate that the accuracy is greater than 95%.

In the proposed work Syntactic features are extracted from Kannada articles to understand the author's writing style. The basic concepts of POS tagging and the different methodology to implement it is discussed in [23]. Authors have served the complete POS tagging information for beginners to carry out research in this domain. They concluded that deep learning algorithms are more powerful compared to traditional methods for the English language.

The authors used deep learning methodologies [24] like RNN and LSTM to assign POS tags to the annotated Kannada words and achieved 81% accuracy. The limitation of this work is in getting the clear dataset in the required format since the same words are spoken and written in different ways due to this one word can have different inflections. This leads to ambiguity in assigning the POS tags.

The authors have explored both the Hidden Markov chain method and conditional random fields algorithms [25] to assign POS tags to the Kannada words. They achieved 79% and 84% for both methods respectively. The model suffers due

to cross domain dataset that is a dataset with different categories.

POS tags are assigned after analyzing each word in the text, authors have demonstrated POS tagging [26] using machine learning algorithms and deep learning algorithms. One of the machine learning algorithms SVM outperforms deep learning techniques with 85% accuracy. The lack of a clean dataset is the only demerit in this work.

Markov chain algorithm again proved to be efficient for a small Kannada dataset [27] of 18,000 words. researchers achieved 95% accuracy but their performance declines as the dataset increased.

Sindhi is one of the oldest languages and not much work have been done in the field of text processing [28], authors have designed rule based approach to assigning POS tags and they were able to assign POS tags successfully on 624 words. Performance comparison has not been focused on since there is no state of art models available for this language.

A large volume of data is flowing over the twitter media, and analyzing the data based on their POS tags are experimented with by the researchers [29]. They have used various classification algorithms to efficiently assign the POS tags to the Malay corpus. SVM yields a maximum accuracy of 95% compared to other algorithms. This approach can be implemented for various categories of Malay words.

Authorship attribution becomes a very important issue in today's time due to the increase in identity theft crimes. The text domain is in ranges from science, art, to philosophy-related texts. It is observed from the survey that, feature extraction plays a huge role in finding the source of a text (Lexical, Semantic and Syntactic features). Language Models (word/ character) and vocabulary of an author also are important parameters in performing the AA task. Few researchers have experimented on both Instance and Profile-based approaches for both global languages and a few Indian local Languages which include short and long texts. For the majority of the AA tasks, the deep learning technique [23] proved to be efficient but can't be claimed as a standardized technique since ML algorithms also outperformed well for other data samples.

There is a research gap in this domain for the Kannada language and this can be used as an opportunity by the interested researchers.

### III. METHODOLOGY

The overall description of the proposed work is given below:

- 1) Let A be the author set  $\{a_i, a_{i+1}, \dots, a_n\}$ .
- 2) Let D be the document set  $\{d_i, d_{i+1}, \dots, d_n\}$  written by the author  $a_i$ , such that  $d_i \in a_i$ .
- 3) Let S be the sentences in a document  $d_i$   $S = \{s_i, s_{i+1}, \dots, s_n\}$  such that  $s_i \in d_i$ .
- 4) Let T be the set of POS tags  $T = \{t_i, t_{i+1}, \dots, t_n\}$  for a sentence created using the CRF algorithm such that  $T \in s_i$ .
- 5) The model extracts semantic features using N-gram technique as  $\{t_1, t_2, t_3\} \dots \{t_i, t_{i-1}, t_n\}$  where it is a POS tag  $\in T$ .



6) Let  $A_t$  be the anonymous text during the testing phase, the authorship model extracts the POS features of the  $A_t$  and compares them with the extracted features and predicts the probability of a true author.

#### A. Dataset Source and Collection

The primary source of the dataset is the internet, from that we selected articles from the Kannada blogs/ websites, e-articles, e-books from Kannada Sahitya Paritshath which is a government-based website and also from other popular Kannada websites. Few authors have contributed their articles based on the request. Totally 500 documents from 50 authors have been considered. It is a cross-domain dataset that each author has written articles on various categories which include: Life skills, philosophy, folk, children's stories, politics and sports. The proposed work is implemented based on the POS and sequence model. The AA is based on the POS tagging to extract the hidden semantic meaning of the text, since there is no open access POS tagging application available on the internet, we created a POS model [24], and the summary of the overall implementation is briefed below:

- Preprocess the dataset by tokenizing the documents (articles/stories/poems) of each author and creating an array of sentences.
- Pass the tokens to a POS tagger developed using the CRF model [25,26]. For the sake of better results, the number of parts of speech is reduced and more generalized, then run the bag of words code into the POS tagger and assign a tag to each word.
- Created embedding vectors, since the model understands only the numbers, not the text.
- Designed a sequential neural network with two LSTM layers, dense layers, activation and dropout layers.
- Train and test the model by running each document sentence-wise.

POS tags are assigned to all the tokens of the author's samples, this gives information about the semantic structure of the language that is the author's usual way of using words to form a sentence. This knowledge helps the model to understand the context of the Kannada words used in a sentence. A word can have different meanings so different POS tags [27,28], the tagger in this case will help a reader to understand the correct meaning of the words based on the tags. The proposed work uses a CRF classifier for assigning the POS tags to the Kannada tokens. The overall implementation of POS tagging in the proposed work is given in Fig. 2.

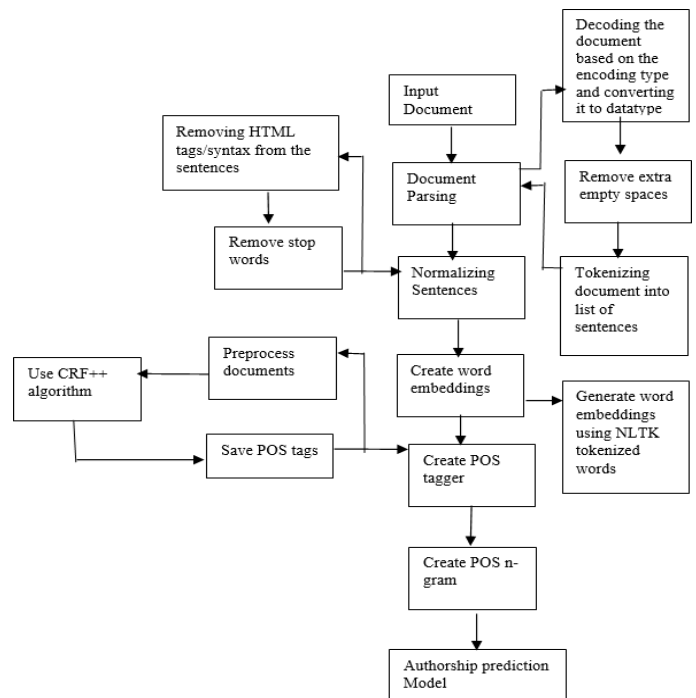


Fig. 2. Working Flow of POS Authorship Model.

#### B. Document Parsing

Document parsing, the first stage consists of three distinct steps.

- Parsing the text document: To build a POS tagging, a separate dataset prepared by the International Institute of Information Technology (IIIT), Hyderabad (IIIT Hyderabad LTRCMT-NLP Lab) is used for training and testing the POS model and Kannada scripts are encoded in the UTF-8 format. Each document will have to be read in the same format to preserve the text. PDF files were not used because no existing tool parsed PDF files with Kannada text in them. The text documents parsed act as the raw source on which various pre-processing steps are carried out:
- Data cleaning: The raw text extracted from the parsing stage is analyzed to check if there are any English words present in them. This stage is executed using the understanding of the range of Kannada words in the UTF-8 encoded format and regular expression. Attempts to remove the stop words were made but not considered for the final evaluation as they would prove to be important in POS detection.
- Tokenization: The text after the previous stage is tokenized into indexes so it can be used later. This process is carried out using out-of-the-shelf methods provided by the Keras API tokenization method. OOV token is used to make sure non-existent words in the dictionary can be marked during model evaluation. Sample output for the same is shown in Fig. 3.

```
{ '<OOV>': 1, 'ಈ': 2, 'ಎಂದು': 3, 'ಬಂದು': 4, 'ಅ': 5,
12, 'ಅವರು': 13, 'ತನ್ನ': 14, 'ಎಂಬ': 15, 'ಅವನ': 16,
'ಬಂದು': 23, 'ಅವನು': 24, 'ನನಗೆ': 25, 'ಅವಳ': 26, 'ಹೆ
3, 'ಎರಡು': 34, 'ಬೇರೆ': 35, 'ಅವರಿಗೆ': 36, 'ಮನೆಗೆ': 37,
'ನಾವು': 44, 'ಜನ': 45, 'ಸ್ವಲ್ಪ': 46, 'ಅಲ್ಲಿ': 47, 'ಎಲ್ಲಾ':
'ಹೆಚ್ಚು': 55, 'ಸರಕಾರ': 56, 'ಕೈ': 57, 'ಬಂದ': 58, 'ಹೊಸ
```

Fig. 3. Sample Tokenized Kannada Words from the Article.

### C. Normalize Sentences

Once the data is parsed and cleaned for POS tagging. Normalizing the sentences on the other hand is used to preprocess the dataset used to train the POS model [11]. The dataset used for POS training comes in HTML format and thus, HTML tags (start tags, end tags, new line and parentheses) are removed to bring the data into raw text format. The raw text is then preprocessed to create a tag set where an index and a Part of Speech are attached to each word in a sentence. The first word of the sentence has index one and the index keeps increasing until a delimiter is found. The first word of the next sentence starts with index 1 again. This tag set contains complex parts of speech attached to each word. To reduce the number of classes for the classification model, we reduce the part of speech to its roots (11 categories: noun, verb, pronoun, intensifier, conjunction, adjective, demonstrative, quantifier, adverb, particle and punctuation). Table I shows the corresponding labels used in the datasets for various POS tags and Fig. 4 shows the output of the POS tagger.

```
[[1, 'ಅಹಂಭಾವವನ್ನು', 'N_NN'], [2, 'ಬಿಟ್ಟು', 'V_VM_VNF',
ಲಿಯಿರಿ', 'V_VM_VF'], [6, '.', 'RD_PUNC'], [1, 'ದೇವಿ',
N'], [5, 'ಪರವಾಗಿ', 'N_NN'], [6, 'ನಾನು', 'PR_PRP'],
C'], [1, 'ಇನ್ನೂ', 'JJ'], [2, 'ಮುಂದೆ', 'N_NN'], [3, '
ಡುವುದಿಲ್ಲ', 'V_VM_VF'], [7, '.', 'RD_PUNC'], [1, 'ಬಿ
ಲ್ಲಿ', 'N_NN'], [5, 'ಇದೋ', 'V_VM_VNF'], [6, 'ಸೋಮ
ತ್ತು', 'N_NN'], [10, 'ಅರಮನೆಯ', 'N_NN'], [11, 'ಪಕ್ಕದ
'ಹಾಕಿಸು', 'V_VM_VF'], [15, '.', 'RD_PUNC'], [1, 'ಸ
```

Fig. 4. POS Tagging for the Tokenized Words.

### D. Creating the embeddings

The identified tokens are used to create a word embedding matrix using the off-the-shelf Language Tokenizer [12]. This produces a 1\*400 vector for each word and an embedding matrix is created by stacking the vectors of each word based on their index from the tokenizer. Language\_tokenizer( )

returns two vectors for certain words as it recognizes the root word and the suffix as two different words. The root word is assumed to have the maximum importance and is retained while the vector for the suffix is discarded. This gives a 27046\*400 dimensions word embedding for the entire dictionary which is used as the first layer in the Sequence model.

TABLE I. KANNADA POS TAGS

Sl. No	POS tags in Kannada	Label
1	ನಾಮಪದ Noun	NP-Noun
2.	ಸರ್ವನಾಮ Pronoun	PR
3.	ಕ್ರಿಯಾಪದ MainVerb	V__VM__VF
4.	ತೀವ್ರಗೊಳಿಸುವಿಕೆ intensifier	RD__INTF
5.	ಸಂಯೋಗ conjunction	CC__CCS
6	ವಿಶೇಷಣ adjective	JJ
7	ಪ್ರದರ್ಶಕ Demonstrative	DM DMD
8	ಪರಿಮಾಣಕಾರಕ Quantifier	QT__QTC
9	ಕ್ರಿಯಾವಿಶೇಷಣ adverb	RB
10	ಕಣ Particle	RP_RPD
11.	ವಿರಾಮಚಿಹ್ನೆ Punctuation	RD_PUNC

### E. Create POS Tagging

The tag set generated in the normalize sentences section is used here. Each word in the POS tag set is processed to contain some features to use in the classification algorithm. The features considered are:

- The word itself.
- The length of the word.
- First 4 letters of the word.
- First 3 letters of the word.
- First 2 letters of the word.
- Last 4 letters of the word.
- Last 3 letters of the word.
- Last 2 letters of the word.
- Is the word a punctuation..
- Surrounding information
  - If the word is not the first word, the above mentioned features of the previous word.
  - If the word is not the second word, the features of the word that are two positions behind.
  - If the word is not the last word, above mentioned features of the next word.
  - If the word is not the last second word, the features of the word are two positions ahead.

The above features combined with N-gram predict the POS tag for a word based on the POS of the previous and the next word of that given word.

The performance of the POS tagger is given in Table II. With the proposed POS model, 91% accuracy is achieved.

TABLE II. PERFORMANCE OF THE PROPOSED POS MODEL

POS tags	Precision	Recall	F1 Score	Support
N	0.843	0.956	0.896	614
V	0.955	0.922	0.938	502
RB	0.444	0.316	0.369	38
CC	0.901	0.839	0.869	87
RD	1.000	0.997	0.998	317
JJ	0.806	0.532	0.641	47
DM	0.927	0.905	0.916	42
RP	0.647	0.333	0.440	33
PR	0.964	0.931	0.948	350
PSP	0.333	1.000	0.500	1
QT	0.913	0.808	0.857	26
Micro average	0.910	0.911	0.910	2057
Macro average	0.794	0.776	0.761	2057
Weighted average	0.909	0.911	0.907	2057
Final F1 score on the test dataset	0.9066			
Accuracy of the test dataset	0.9101			

The proposed POS model's performance is compared with those of other models identified during the survey. We were able to achieve a decent accuracy of 91% accuracy and a 90% F1 score since the model employs a clear dataset. The performance of the POS model is seen in Table III. The CRF model is one of the top models for assigning POS tags for Kannada words, according to the survey.

TABLE III. PERFORMANCE ANALYSIS OF VARIOUS POS MODEL

Reference No	Method employed	Accuracy in %
[24]	RNN+LSTM	81
[25]	Hidden Markov model CRF	79 84
[26]	SVM	85
[27]	Markov chain	95
[29]	SVM	95
<b>Proposed model</b>		<b>91</b>

#### F. Authorship model using Bi-LSTM

Once the author's dataset is prepared and the POS model assigns the tags for each token in the dataset, the next stage is, running the proposed authorship model. It is a classification problem and Bi-LSTM is employed to perform the AA task. Bi-LSTM model [13] reads the author's text in both directions and understands the syntactic features, especially how the sentence is formed using the POS tags. The Bi-LSTM architecture used for the proposed work is shown in Fig. 5 and the simplified architecture is shown in Fig. 6.

- Layer 1 composing 15 Bi-Directional LSTM Units: A Bidirectional LSTM, or Bi-LSTM, is a sequence processing model that consists of two LSTMs [10]: one taking the input in a forward direction, and the other in a backward direction. Bi-LSTMs effectively increase the amount of information available to the network, improving the context available to the algorithm (for example, knowing what words immediately follow and precede a word in a sentence).
- Layer 2 stacked on top of layer 1 consisting of 15 Bi-Directional LSTM Units.
- Batch Normalization for the sequence: It is a process to make neural networks faster and more stable by adding extra layers to a deep neural network. The new layer performs the standardizing and normalizing operations on the input of a layer coming from a previous layer.
- A densely connected neural network layer.
- ReLU activation: A linear function that will output the input directly if it is positive, otherwise, it will output zero.
- 64 Dense units with the 'ReLU' activation and a dropout of 0.5 32 Dense units with the 'ReLU' activation and a dropout of 0.5.
- Dropout: Dropout is a technique used to prevent a model from overfitting. Dropout works by randomly setting the outgoing edges of hidden units (neurons that make up hidden layers) to 0 at each update of the training phase.
- Sigmoid activation: A sigmoid function is a bounded, differentiable, real function that is defined for all real input values and has a non-negative derivative at each point and exactly one inflection point.

The model is compiled with the following hyperparameters:

- Loss: Binary cross entropy compares each predicted probability to actual class output, which can be 0 or 1. It then calculates the score that penalizes the probabilities based on the distance from the expected value. That means how close or far from the actual value.
- Optimizer: Stochastic Gradient Descent with a learning rate of 0.001 and momentum of 0.9. It attempts to find the global minimum by adjusting the configuration of the network after each training point. Instead of decreasing the error, or finding the gradient, for the entire data set, this method merely decreases the error by approximating the gradient for a randomly selected batch (which may be as small as a single training sample). In practice, random selection is achieved by randomly shuffling the dataset and working through batches in a stepwise fashion.
- Prediction: Each of the text documents is preprocessed similarly during the training and validation sets.

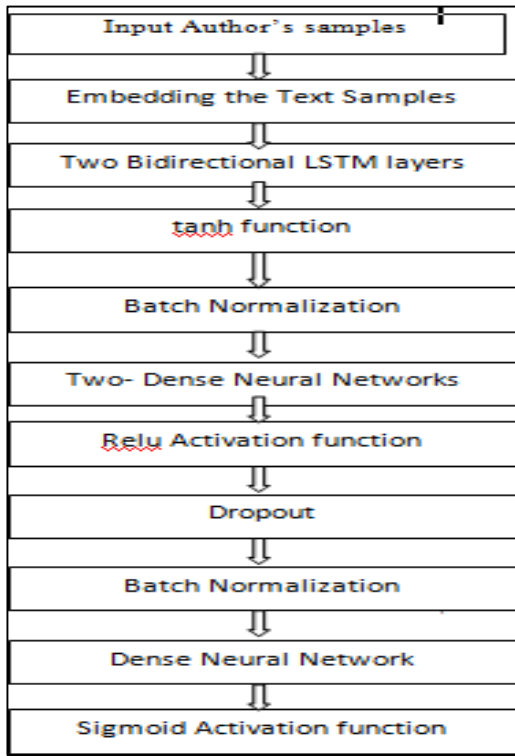


Fig. 5. Overall Bi- LSTM Architecture of the Proposed Model.

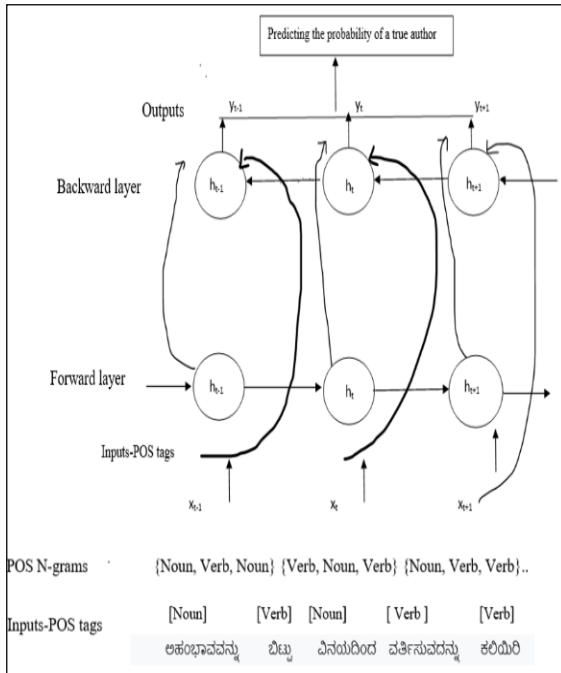


Fig. 6. The Architecture of the AA Model using Bi-LSTM.

All the documents with the labeled author's names are trained but testing is done in a different approach, that is 10 datasets are created such that, each dataset comprises articles written by five different authors so a total of 50 author's 10 documents are collected. Table III shows the dataset preparation for the proposed model. Finally, 500 articles are in the dataset, it is split into 70:20:10 for training, testing and

validation respectively. The model is trained using the N-gram approach by extracting the POS tag features of the authors. During the testing phase, if author 1's document is passed, then Model 1 is supposed to respond positively to the statements of author 1 and should predict 0 for authors 2,3,4 and 5's documents. Similarly, model 10 contains articles written by the authors 46,47,48,49 and 50. After predicting the author for each sentence in each document, the ratio of positive to negative statements is calculated. We used the 1v5 method for testing rather than having 5 neurons that associate a document to an author with a certain probability since this model produced more promising results.

#### IV. RESULTS AND DISCUSSION

The performance of 5 author sets of 10 models is tabulated based on the N-grams of POS tags. One of the metrics to measure the performance of the proposed model is the count of positive and negative statements.

The overall loss rate of all the dataset models is shown in Fig. 7, the loss rate is reduced after several epochs which indicates that the model is learning the writing style of the authors in a better way. After testing the author's samples with the different combinations, we obtained the accuracy of all ten author sets and tabulated them in Table IV and Fig. 8 to indicate the same.

We also observed that for a few articles authors are mispredicted. The performance is mainly depending on the efficiency of the POS tagger [12]. POS tagger is working fine with the training set but has average performance for the authorship dataset. The proposed work uses cross domain authorship that is an author can write political as well as scientific articles since the writing style differs, it has an impact on the accuracy and the N-gram technique extracting the POS tags is not sufficient to extract the writing style, N-gram combined with other stylometry features may work better. The performance of the model deteriorates when the anonymous text document is tested against the articles of all 50 authors. The accuracy was 10%-12% so a 1v5 approach is used. In the 1v5 approach, each time the dataset model contains different authors' articles.

TABLE IV. AUTHOR WISE ACCURACY

Dataset Models	Accuracy
Model-1	77%
Model-2.	78%
Model-3	77%
Model-4	78%
Model-5	81%
Model-6	77%
Model-7	81%
Model-8	78%
Model-9	72%
Model-10	79%
Average Accuracy	77.8%

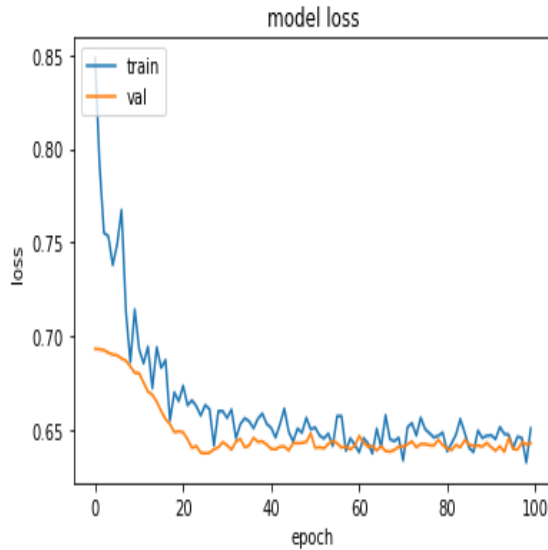


Fig. 7. Loss Rate.

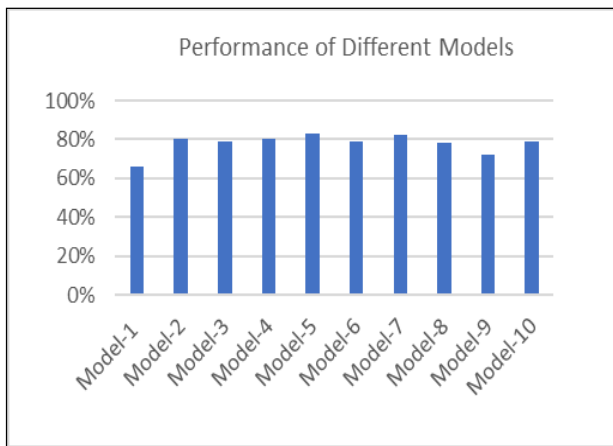


Fig. 8. Accuracy of All the Dataset Models.

Using LSTM one more approach called the sequence model is developed. This experimental model was used without convincing results. The text source created at the end of step 3 as shown in Fig. 5 is used for this model. After splitting the dataset as train, test and validation sets like the BI-LSTM model. All 50 authors were used for this model and the multi-class classification strategy was employed here. The prediction vector was in the form of a one-hot encoded vector where the index of the corresponding author was marked as 1 and the rest as 0. The sets from the previous stage are converted to indices using the tokenization from stage 4 of the Data Processing module. Both sentence level and document level sequences were considered with lengths 15 and 2500 respectively but this model yields overall accuracy of 15%.

#### A. Performance Comparison of Different Techniques

To assess the performance of our model, we have compared it with other AA models developed for other languages since no models are available for the Kannada languages. Table V shows the accuracy obtained by other AA

models. We observed that deep learning based models performed well with more than 90% accuracy, but we obtained 77.8% accuracy for the proposed work. Also, classification algorithms perform better in some cases. The reasons for the moderate performance of the proposed model are identified and listed below:

- The POS model is tested and trained successfully on a labeled dataset, but when it is used on a real authors dataset, it performs moderately since certain Kannada words have several meanings, which causes ambiguity.
- In the proposed work a cross-domain dataset is considered which has a variety of categories like life skills, science and health, sports, etc. It is common in foreign languages but is new to the Kannada language.
- A good dataset of more than 10,000 articles improves the model's performance, however building such a huge Kannada dataset is a challenging one.

TABLE V. PERFORMANCE COMPARISON OF AA MODELS

Reference No	Technique used	Accuracy / F1 Score in %
[9]	Artificial Neural Network	75.46
[10]	N-gram	61.1
[11]	Life Like Network Automata using Life-Like rules	70-75
[12]	SVM algorithm with linear kernel	68
[14]	SVM	96
	Deep learning	93
[16]	Convolutional Neural Network	98
[17]	BERT model	90
[18]	Neural Networks	97.2
[19]	Doc2vec Reduction model	99
[20]	Recurring Neural Networks	78
[21]	Convolutional Neural Network	72.59
[22]	SVM	95
<b>Proposed Model</b>	<b>Bi-LSTM</b>	<b>77.8</b>

#### B. Directions for Future Work

More powerful style markers can be introduced which can be used to classify a wide range of authors [30]. Right now, this work can classify 50 authors. The next step can be the development of general rules that should apply to almost every author. A near perfect classification can be achieved by training a meta-learner that will use both neural networks and decision trees, this is required because neural networks consider different sets of features than decision trees.

A combination of different sets of features may be tried to see if there exists a set that can be used by all learning algorithms. Also, feature extraction can be improved by using more powerful natural language tools. Features selection is an important criterion and the graph-based neural networks [14] can be used in the future where the network will itself select the features which contribute the most to the output.

## V. CONCLUSION

The model currently stands with an accuracy of 77.8%. According to our knowledge, the proposed model is a novel technique and a challenging one in recognizing the writing style of a Kannada author. The proposed work aims to increase the accuracy by tweaking the model and if possible, implementing the same model by extracting features other than syntactic. It is understood that lexical features, word length and sentence length do not often have enough descriptive power for any model to assign a document to an author with a sense of certainty. We aim to pursue further research in detail to pick only those features that will yield good results. We believe semantics is one of those.

## REFERENCES

- [1] Shreyashree, S., Sunagar, P., Rajarajeswari, S. and Kanavalli, A, "A Literature Review on Bidirectional Encoder Representations from Transformers," Lecture Notes in Networks and Systems, Springer, Singapore, vol 336, pp. 305-320, Jan 2022.
- [2] Pramod Sunagar and Anita Kanavalli, "A Hybrid RNN based Deep Learning Approach for Text Classification," International Journal of Advanced Computer Science and Applications(IJACSA), Vol13(6), pp. 289-295, July 2022.
- [3] B.V.Dhandra and M.B.Vijayalaxmi, "A Novel Approach to Text Dependent Writer Identification Of Kannada Handwriting", Procedia, CS, Volume 49, pp. 33-41, 2015.
- [4] Praveen Bangarimath and DeepaBendigeri, "Writer Identification using Texture Features in Kannada Handwritten Documents", IICA Proceedings on National Conference on Electronics, Signals and Communication, Vol 6 13, pp.5-8, 2018.
- [5] Fakhraddin Alwajih, Eman Badrand and Sherif Abdou, "Transformer-based Models for Arabic Online Handwriting Recognition," International Journal of Advanced Computer Science and Applications (IJACSA), Vol. 13, No. 5, pp. 898-905, 2022.
- [6] Chandrika C.P, Kallimani, J.S, "Authorship Attribution for Kannada Text Using Profile Based," Proceedings of the 2nd International Conference on Recent Trends in Machine Learning, IoT, Smart Cities and Applications, Lecture Notes in Networks and Systems, Springer, Singapore, vol 237, pp. 679-688, Jan 2022.
- [7] Chandrika C.P and Kallimani, J.S, "Instance Based Authorship Attribution for Kannada Text Using Amalgamation of Character and Word N-grams Technique", Distributed Computing and Optimization Techniques, Lecture Notes in Electrical Engineering, Springer Singapore, vol 903, pp 547-557, Aug 2022.
- [8] Efstathios Stamatatos. "A Survey of Modern Authorship Attribution Methods," Journal of the American Society for Information Science and Technology, vol 60, pp 538-556, March 2009.
- [9] Mohammad Al-Sarem, Abdullah Alsaeedi, and Faisal Saeed, "A Deep Learning-based Artificial Neural Network Method for Instance-based Arabic Language Authorship Attribution", International Journal of Advances in Soft computing and its Applications, ISSN 2074-852, Vol. 12, pp. 1-14, Dec 2020.
- [10] Yaakov HaCohen-Kerner, Daniel Miller, Yair Yigal, and Elyashiv Shayovitz, "Cross-domain Authorship Attribution: Author Identification using Char Sequences, Word Uni-grams, and POS-tags Features," Notebook for PAN competition at CLEF 2018.
- [11] Machicao J, Corréa EA Jr, Miranda GHB, Amancio DR and Bruno OM, "Authorship attribution based on Life-Like Network Automata", PLoS One, Vol:13(3), DOI: 10.1371/journal.pone.0193703, March 2018.
- [12] Andrea Bacciu, Massimo La Morgia, Alessandro Mei, Eugenio Nerio Nemmi, Valerio Neri, and Julinda Stefa, "Cross-Domain Authorship Attribution Combining Instance-Based and Profile-Based Features," A notebook for PAN competition at CLEF 2019.
- [13] C. NamrataMahender, Ramesh Ram Naik and Maheshkumar Bhujangrao Landge, "Author Identification for Marathi Language," Advances in Science, Technology and Engineering Systems Journal, India, ISSN: 2415-6698, Vol. 5, No. 2, pp. 432-440.
- [14] Aleksandr Romanov, Anna Kurtukova , Alexander Shelupanov, Anastasia Fedotova and Valery Goncharov, "Authorship Identification of a Russian-Language Text Using Support Vector Machine and Deep Neural Networks", Future Internet 2021, Vol: 13, pp. 1-16, Dec 2020.
- [15] Aisha Khatun, Anisur Rahman, Md. Saiful Islam and Marium-E-Jannat, "Authorship Attribution in Bangla literature using Character-level CNN," 22nd International Conference on Computer and Information Technology (ICCIT), pp 1-5, 2019.
- [16] Georgios Barlas and Efstathios Stamatatos, "Cross-Domain Authorship Attribution Using Pre-trained Language Models," Artificial Intelligence Applications and Innovations, IFIP Advances in Information and Communication Technology, Vol 583, DOI:https://doi.org/10.1007/978-3-030-49161-1\_22, pp 255-266, May 2020.
- [17] S. H. H. Ding, B. C. M. Fung, F. Iqbal and W. K. Cheung, "Learning Stylometric Representations for Authorship Analysis," IEEE Transactions on Cybernetics, DOI: 10.1109/TCYB.2017.2766189, vol. 49, no. 1, pp. 107-121, 2019.
- [18] Michael Tschuggnall, Benjamin Muraier and Gunther Specht, "Reduce & Attribute Two-Step Authorship Attribution for Large-Scale Problems," Proceedings of the 23rd Conference on Computational Natural Language Learning, Hong Kong, China, pp. 951-960, Nov 2019.
- [19] Fereshteh Jafariakinabad, Sansiri Tampradab and Kien A. Hua, "Syntactic Neural Model for Authorship Attribution," The Thirty-Third International FLAIRS Conference (FLAIRS-33), pp.1-6, 2020, May 2020.
- [20] Ehsan Reisi1 and Hassan Mahboob Farimani, "Authorship Attribution in Historical and Literary Texts by A Deep Learning Classifier," Journal Of Applied Intelligent Systems & Information Sciences, Vol 1. Issue 2, pp. 118-127, December 2020.
- [21] Suhad A. Yousif, Zainab N. Sultani, Venus W. Samawi, "Utilizing Arabic WordNet Relations in ArabicText Classification: New Feature Selection Methods," IAENG International Journal of Computer Science, Vol 46:4, 2019.
- [22] Qinghe Zheng, Xinyu Tian, Mingqiang Yang and Huake Su, "The Email Author Identification System Based on Support Vector Machine (SVM) and Analytic Hierarchy Process (AHP)", IAENG International Journal of Computer Science, vol 46:2, pp.1-14, May 2019.
- [23] Chiche, A and Yitagesu, B, "Part of speech tagging: a systematic review of deep learning and machine learning approaches," J Big Data, Vol9(10), United Kingdom, pp.2-25, Jan 2022.
- [24] Rajani Shree, M., Shambhavi, B.R , " POS Tagger Model for South Indian Language Using a Deep Learning Approach", Lecture Notes in Electrical Engineering, Springer, Singapore, vol 828, pp.26-30, Jan 2022.
- [25] Shambhavi, Ravi and P Ramakanth, "Kannada Part-Of-Speech Tagging with Probabilistic Classifiers," International Journal of Computer Applications, Vol 48. pp.26-30, June 2012.
- [26] Shriya Atmakuri, Bhavya Shahi, Ashwath Rao B and Muralikrishna SN, "A comparison of features for POS tagging in Kannada," International Journal of Engineering & Technology, vol 7, pp.2418-2421, 2018.
- [27] Saritha Shetty and Savitha Shetty, "Text pre-processing and parts of speech tagging for Kannada language," Journal of Xi'an University of Architecture & Technology, vol 11, pp 1286- 1291,2020.
- [28] Irum Naz Sodhar, Abdul Hafeez Buller, Suriani Sulaiman and Anam Naz Sodhar, "Word by Word Labelling of Romanized Sindhi Text by using Online Python Tool" International Journal of Advanced Computer Science and Applications(IJACSA), vol 13(8), pp.262-267, 2022.
- [29] Siti Noor Allia, Noor Ariffin and Sabrina Tiun, "Improved POS Tagging Model for Malay Twitter Data based on Machine Learning Algorithm", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 13, No. 7, pp 229-234, 2022.
- [30] Patrick Juola, "Future Trends in Authorship Attribution," IFIP International Federation for Information Processing, Advances in Digital Forensics III, Volume 242, pp. 119-132.



# Flood Prediction using Deep Learning Models

Muhammad Hafizi Mohd Ali, Siti Azirah Asmai\*, Z. Zainal Abidin, Zuraida Abal Abas, Nurul A. Emran  
Centre for Advanced Computing Technology (C-ACT), Fakulti Teknologi Maklumat Dan Komunikasi  
Universiti Teknikal Malaysia Melaka, Hang Tuah Jaya, 76100 Durian Tunggal, Melaka, Malaysia

**Abstract**—Deep learning has recently appeared as one of the best reliable approaches for forecasting time series. Even though there are numerous data-driven models for flood prediction, most studies focus on prediction using a single flood variable. The creation of various data-driven models may require unfeasible computing resources when estimating multiple flood variables. Furthermore, the trends of several flood variables can only be revealed by analysing long-term historical observations, which conventional data-driven models do not adequately support. This study proposed a time series model with layer normalization and Leaky ReLU activation function in multivariable long-term short memory (LSTM), bidirectional long-term short memory (BI-LSTM) and deep recurrent neural network (DRNN). The proposed models were trained and evaluated by using the sensory historical data of river water level and rainfall in the east coast state of Malaysia. It were then, compared to the other six deep learning models. In terms of prediction accuracy, the experimental results also demonstrated that the deep recurrent neural network model with layer normalization and Leaky ReLU activation function performed better than other models.

**Keywords**—Deep learning; recurrent neural network; long short-term memory; flood prediction; layer normalization

## I. INTRODUCTION

Due to its impact on daily life, flooding is one of the most pressing issues that Malaysia has been dealing with recently. Floods are a type of natural geohazard that typically occur because of consistently heavy rain. This natural phenomenon causes massive damage to the country's property and Gross Domestic Product (GDP). According to Ashizawa et al. [1], the entire GDP of Japan impacted by flood damage is at least 1% of the overall GDP of the nation. Tiggeloven et al. [2] stated that the top 15 countries, such as India, Bangladesh, China, and others, are vulnerable to flood occurrence at the present day and could be worst if no action is taken. Indeed, floods can cause a massive amount of money to repair the damage. Hence, flood occurrence can affect every country, including Malaysia. Shaari et al., [3] stated that from 2006 to 2010, there was nearly 1 million USD damage caused by floods which affected the nation's economic growth. There are many classifications of floods namely coastal floods, flash floods, ponding (or pluvial flooding), and river (or fluvial) floods [4]. Floods often occur, especially in Southeast Asia, including our country, Malaysia. The general types of flooding in Malaysia include riverbank overflow, flash floods, high tides [5], and monsoon floods [6].

Floods are classified as natural disasters in Malaysia due to the monsoon season. In Malaysia, there are two distinct monsoon seasons: the Northeast Monsoon, which occurs from November to March, and the Southwest Monsoon, which occurs from late May to September. The Northeast Monsoon

can bring heavy rainfall. Due to the extensive network of rivers connecting several Malaysian states and the country's poor drainage, floods are expected in Malaysia during the monsoon season, especially on the West Coast and in Borneo. The populous region is flooded as a result of the river level rising significantly as a consequence of these strong rains. As a result, people are compelled to temporarily relocate to several relief. Floods halt economic progress since crops and animals are destroyed. Romali et al, [7] stated that Malaysian financial losses are estimated at nearly MYR 915 million annually on an average due to floods.

According to Zerara [8], time series is a statistical method that can be applied in a broad range of longitudinal research designs. Typically, this time series design involves a single subject that is measured repeatedly at regular intervals over a large number of observations. Time series forecasting aims to predict an outcome based on the collection of historical data that can be used to build a quantitative model that explains the variables under consideration [9]. For many years, time series forecasting has been an important research domain in meteorology [10], biology [11], and econometrics [12]. Generally, time series can have four characteristics: trends, seasonality, cycles, and noise [13]. Time series forecasting algorithms perform well with data that includes a time dimension and one or more properties [14].

Time series have been frequently utilized in flood forecasting and have shown excellent results for the global community [15]. Furthermore, according to Shen et al. [16], modern time series can be combined with deep learning models including Recurrent Neural Network (RNN), Artificial Neural Network (ANN), Long Short Term Memory (LSTM), and other models. However, the existing flood forecasting methods, for example, frequency analysis, rational method, and empirical formula, are not deemed suitable for a wide area. Those methods can only cover a small river flow area [17]. For example, Faruq et al. [18] used an LSTM model to predict the flood by using a Klang River lever dataset. Another study [19] used the ANN model to predict floods by using the Kelantan river lever and rainfall dataset in separate models. The LSTM model performs exceptionally well when the modelling has a large amount of time series data. Furthermore, according to the literature, the LSTM model outperforms the RNN in the number of previous time steps that can be considered. Additionally, Siami-Namini & Namin, [20] demonstrated that the LSTM model could predict time series much more accurate than the Autoregressive Integrated Moving Average (ARIMA) model in some cases. As a result, the LSTM model appears to be a likely top performer; although the study employs stock time series, it still has parallels to the current study since it focuses on time series forecasting. According to Jaiswal & Das

\*Corresponding Author.

[21], the ANN model works best with nonlinear problems whereas Šiljić Tomić et [22] stated that the ANN model can work with both nonlinear and linear problems. According to Y.f. Zhang et al. [23], despite being the most significant advantage of time series, long-term dependencies remain a considerable challenge. Besides, one of the most common drawbacks of the LSTM model and BI-LSTM model is the high computational cost during training procedures [24], where potential time series forecasting in floods has yet to be unfolded. Despite advances in developing models based on RNN, these models remain challenging to scale to long data sequences. Dhunny et al. [25] have proven that an ANN model can predict the flood water level well within 24 hours ahead of time by using the data from rainfall and present river level data. In this study, the flood was predicted for one day ahead.

The rest of this paper is structured as follows. Section II describes the related work and literature for this study. Section III presents the proposed models for flood prediction. Section IV covers the experimental procedure for the case study whereas the findings are discussed in Section V. Finally, Section VI concludes the paper, and a discussion of future works is included in Section VII.

## II. RELATED WORK

Artificial neural networks, often known as deep learning, are machine learning algorithms that have been influenced by the structure and operation of the human brain. Deep learning has dominated many uses and has proven to be superior to traditional machine learning algorithms because it can produce faster and more accurate results [26].

Deep learning enables computational models comprising multiple hidden layers of artificial neural networks with multiple linear and nonlinear transformations that learn the data representations with multiple abstraction levels [27]. In deep learning, multiple layers of nonlinear processing units perform feature extraction transformation in the deep learning model. Every layer in the previous layer input is used as its output and it is used in both supervised and unsupervised methods for classification problems and pattern analysis problems [28]. The characteristic of neural networks can be seen in Fig. 1.

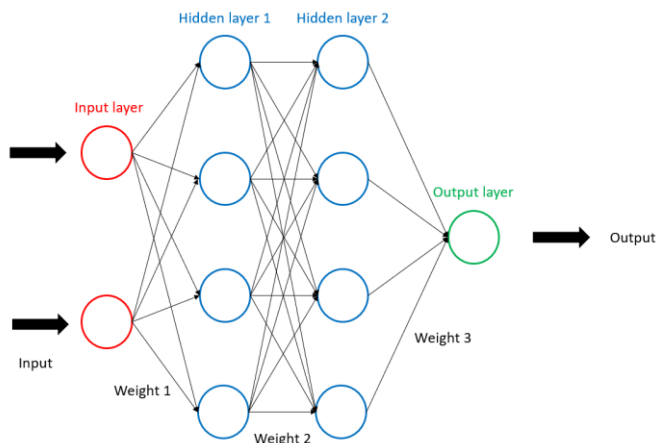


Fig. 1. A Characteristic of Neural Network.

The human brain structure inspires the neural network architecture. Our brains can be trained to recognise patterns and classify various types of information. The possibility of detecting and displaying the correct answer increases with each layer of a neural network, which may be thought of as a kind of filter that operates from coarse to fine.

The neural network can begin to identify trends across the many samples it processes and classify data based on their similarities by using several layers of functions to decompose unstructured data into data points and information that a computer can use.

After processing a large number of structured data training samples, the algorithm has created a model of which elements in data and their relationships must be taken into account when determining whether structured data is present or not. The neural network compares new data points to its model based on all previous evaluations when evaluating new data. The model is then used to determine whether the data contains specific data.

The layers of functions that are present between the input and output in this example serve as a representation of deep learning. The interaction across layers is marginally enhanced in the following Fig. 2, however, the connections between nodes or artificial neurons might vary significantly.

### A. Long Short-Term Memory (LSTM)

Long Short-Term Memory (LSTM) networks can be considered as recurrent neural networks that are modified to improve from the RNN model with memory recall function. The LSTM classifier is ideal for processing, classifying, and forecasting time series with unknown time lags. Backpropagation is applied when training the model. There are three gates in the LSTM model as shown in Fig. 3.

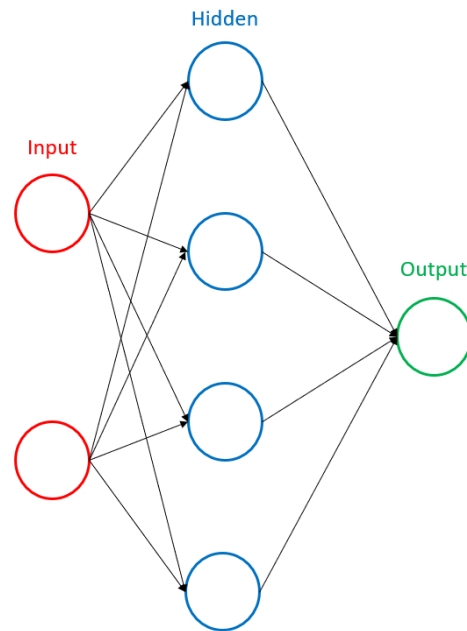


Fig. 2. Interaction between Layers.

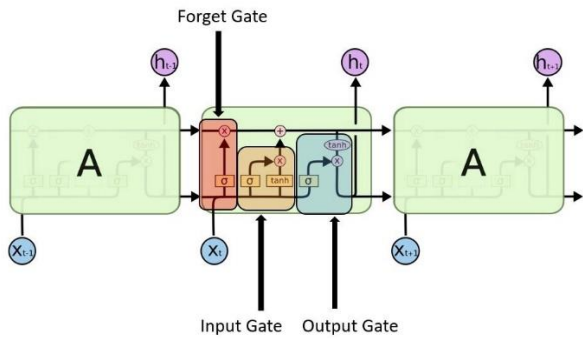


Fig. 3. LSTM Gates [29].

Forget gate - it analyses the previous state ( $h_{t-1}$ ) and the input content ( $x_t$ ) and returns the value ranging from 0 to 1 for each number in the cell state  $C_{t-1}$  by deciding through the sigmoid function. The forget gate equation of [30] can be referred to as (1).

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad (1)$$

Input gate - values from the input are applied to adjust the memory. The sigmoid function determines the value between 0 and 1, allowing through. The tanh function gives weightage to the values and then passes it, specifying their significant level ranging between -1 and 1. The input gate equation of [30] can be referred to as (2).

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C)$$

$$C_t = f_t \times C_{t-1} + i_t \times \tilde{C}_t \quad (2)$$

Output gate - to determine the output, the memory and input of the block are applied. The sigmoid function determines the value between 0 and 1, allowing through. The tanh function gives weightage to the values and then passes it, specifying their significant level ranging between -1 to 1 and multiplying it by the sigmoid output. The forget gate equation of [30] can be referred to as (3).

$$o_t = \sigma(W_o [h_{t-1}, x_t] + b_o)$$

$$h_t = o_t * \tanh(C_t) \quad (3)$$

### B. Recurrent Neural Network

A recurrent neural network (RNN) is a simplified feedforward neural network with internal memory as seen in Fig. 4. RNN is recurrent because it works with the same function for each input data, and the current input as output is dependent on the previous computation. The output is produced, replicated, and then returned to the recurrent network. When making a decision, the current input and output might be seen as learned from the past.

By taking advantage of the internal state (memory) that the RNN model has, it can process the input sequence that is different from typical feedforward neural networks. As a result, the RNN model is suitable for speech recognition, unsegmented, or handwriting recognition tasks. However, there is one drawback in the RNN model: a vanishing gradient problem when dealing with long sequence data. The inputs are

completely independent of one another for other neural networks. On the other hand, all the inputs are connected in RNN.

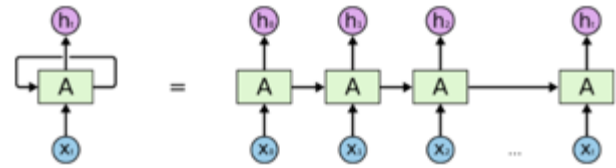


Fig. 4. An Unrolled RNN [29].

The  $X(0)$  is taken from the input sequence at first, then the output  $h(0)$ , with  $X(1)$ , as the input for the next step. The  $h(0)$  and  $X(1)$  are the input for the next step. Similarly,  $h(1)$  from the previous step becomes the input for the next step of  $X(2)$ .

The formula equation of [31] can be referred to as (4) and this is the current state of the equation.

$$h_t = f(h_{t-1}, x_t) \quad (4)$$

Activation Function is applied:

$$h_t = \tanh(W_{hh} h_{t-1} + W_{xh} x_t) \quad (5)$$

The equation of [31] can be referred to as (5), in which  $W$  is the weight,  $h$  is the single hidden vector,  $Whh$  is the weight at the previous hidden state,  $Whx$  is the weight at the current input state, and  $\tanh$  is the activation function that applies a non-linearity squash of the activations to the range between [-1.1].

Output:

$$y_t = W_{hy} h_t \quad (6)$$

The equation of [31] can be referred to as (6), in which the output state is  $Y_t$ . The weight in the output state is  $Why$ .

Flood forecasting is a tool that allows flood control management to predict when local flooding is likely to occur with a high degree of accuracy. The river basin or watershed size can indicate the water levels and flow rates for intervals ranging from a few hours to days ahead; forecasted streamflow and precipitation data are used in the streamflow routing model and rainfall runoff. Flood forecasting can also use precipitation forecasts to expand the available lead time.

Flood forecasting is a crucial component of a flood warning system. The difference between flood forecasting and a flood warning is that flood forecasting produces a set of forecast time that profiles the river levels or flows channel at different locations. In contrast, "flood warning" refers to using forecasts to inform about flood warnings. A popular method applied for flood forecasting is hydrological modelling because this model is a simplified representation of a real-world system [32]. Although this model is good, the downside is that it is a scaling problem that faces a scale area parameter [33].

The existing flood forecasting method cannot be used with a traditional database based on a single source as the main data [34] and it requires a lot of data. With the current technology, flood prediction is more robust, and real-time flood forecasting in the provincial area can be accomplished quickly by utilising the technology of artificial intelligence (AI) and fourth

industrial technology (4IR). An effective real-time flood forecasting model may be helpful for disaster prevention, offering an advanced alert and mitigating the damage from the flood occurrence [35]. Flood forecasting has been improved by utilising deep learning models such as LSTM, RNN, and many others [18]. Many studies have applied a deep learning model in their study to predict flood occurrence and are proven to be an informative and accurate model as shown in Table I. Hence, for the deep learning model, there is always room for improvement with the use of uncertain data like flood data with a nonlinear characteristic. However, there are very few studies that build a deep learning model for flood prediction using multivariate data. In addition, the majority of past research has only used one variable or input as their main source of information to forecast the flood. In addition, the previous research already demonstrates a high level of accuracy. However, the majority of these studies utilize basic deep learning models without any additional characteristics, which can be seen as a discrepancy between the studies. The proposed models aim to get a dependable and more accurate prediction model while reducing the limitations of the previous study by using an additional characteristic as described in Section III. As previous studies are concerned with using a single data as primary data, this study proposed multivariate data as primary data to determine a correlation between several variables simultaneously and a deeper understanding of how the multivariate data relate to real-world scenarios like flood occurrence.

TABLE I. LITERATURE SUMMARY

Models	Title (Author and Year)	Goal	Country
(LSTM) and Radial basis function neural network (RBFNN)	Deep Learning-Based Forecast and Warning of Floods in Klang River, Malaysia [18]	Forecasting the river water level in the Klang River basin, Malaysia.	Malaysia
LSTM and RNN	Application of Long Short-Term Memory (LSTM) neural network for flood forecasting [36]	Proposing an effective approach to flood forecasting based on the data-driven method.	Vietnam
ANN	Flood Prediction through Artificial Neural Networks: A case study in Goslar [37]	Establishing, training and evaluating a neural network for the detection of flood hazards and concrete water levels.	Germany
LSTM	Flood Prediction and Uncertainty Estimation Using Deep Learning [38]	Exploring the deep learning model for predicting gauge height and evaluating the associated uncertainty	United States of America
LSTM	Flash flood forecasting based on long short-term memory networks [39]	Forecasting a model based on (LSTM) for flash flood forecasting.	China
ARIMA and LSTM	Forecasting Economic And	Investigating which forecasting methods	Not stated

	Financial Time Series: ARIMA Vs LSTM [20]	offer the best predictions with the lower forecast errors and higher accuracy of forecasts	
LSTM and TCN	An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modelling [40]	Raising issues of whether these successes of convolutional sequence modelling are confined to specific application domains or whether a broader reconsideration of the association between sequence processing and recurrent networks are in order.	Not stated
LSTM and TCN	Temporal Convolutional Networks Applied to Energy-related Time Series Forecasting [41]	Proposing a TCN-based deep learning model to improve the predictive performance in energy demand forecasting	Spain

### III. PROPOSED MODELS

The proposed enhanced models for flood prediction were formulated to increase the prediction accuracy. In this study two methods in the models were introduced as follows:

#### A. Layer Normalization

Inspired by the results of Batch Normalization, the Layer Normalization method is proposed by normalizing activations along the feature direction rather than the mini-batch direction. Hence, overcoming the disadvantages of batch normalization by eliminating the reliance on batches and making it easier to apply for RNN. Each activation feature is normalized to zero mean and unit variance through Layer Normalization.

In Batch Normalization, the statistics are computed across the batch, as for the spatial dimensions. In contrast, Layer Normalization (LN) computes statistics (mean and variance) across all channels and spatial dimensions. As a result, the statistics are batch independent. This layer was initially designed to handle vectors (mainly the RNN outputs).

Layer Normalization visually comprehends this as shown in Fig. 5:

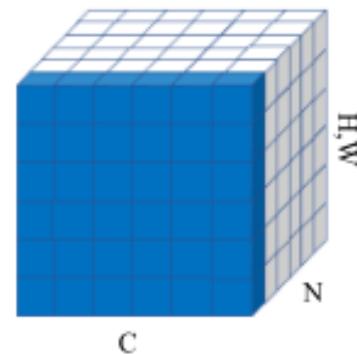


Fig. 5. An Illustration of Layer Normalization [42].

When dealing with vectors with a batch size of NN, the 2D tensors of shape R N times K RNK.

Normalize with the mean and variance of each vector because it does not depend on the batch and its statistics. The normalize equation of [43] can be referred to as (7).

$$\begin{aligned} \mu_n &= \frac{1}{K} \sum_{k=1}^K x_{nk} \\ \sigma_n^2 &= \frac{1}{K} \sum_{k=1}^K (x_{nk} - \mu_n)^2 \\ \hat{x}_{nk} &= \frac{x_{nk} - \mu_n}{\sqrt{\sigma_n^2 + \epsilon}}, \hat{x}_{nk} \in R \end{aligned} \quad (7)$$

$$LN_{\gamma, \beta}(x_n) = \gamma \hat{x}_n + \beta, x_n \in R^K$$

When generalizing to 4D feature map tensors, it takes the mean across all channels and spatial dimensions, as shown below: The equation below based on [43] can be referred to as (8).

$$\begin{aligned} LN(x) &= \gamma \left( \frac{x - \mu(x)}{\sigma(x)} \right) + \beta \\ \mu_n(x) &= \frac{1}{CHW} \sum_{c=1}^C \sum_{h=1}^H \sum_{w=1}^W x_{nchw} \\ \sigma_n(x) &= \sqrt{\frac{1}{CHW} \sum_{c=1}^C \sum_{h=1}^H \sum_{w=1}^W (x_{nchw} - \mu_n(x))^2} \end{aligned} \quad (8)$$

### B. Leaky ReLU

To replace its saturated counterpart of Sigmoid or Tanh, the modern deep learning system employs a non-saturated activation function such as ReLU and Leaky ReLU. It solves the "exploding/vanishing gradient" issue and speeds up convergence.

ReLU reduces the negative component to zero while keeping the positive component. It has the desirable property of being sparse in activations after passing through ReLU. The equation of [44] can be referred to as (9).

$$y_i = \begin{cases} x_i & \text{if } x_i \geq 0 \\ 0 & \text{if } x_i < 0 \end{cases} \quad (9)$$

The gradient-based optimization algorithm will not change the weights of a unit that does not initially activate. Because the gradient is 0 when the unit is inactive, ReLU has a disadvantage during optimization.

If the neurons are not activated at the start of the ReLU, it is possible to end up with a neural network that never learns. The learning rate is slow when training ReLU networks with constant 0 gradients. The equation of [44] can be referred to as (10).

$$y_i = \begin{cases} x_i & \text{if } x_i \geq 0 \\ \frac{x_i}{a_i} & \text{if } x_i < 0 \end{cases} \quad (10)$$

Leaky ReLU adds a slight negative slope to the ReLU to sustain and keep the weight updates alive throughout the propagation process.

The alpha parameter was introduced to address the ReLU's dead neuron issues, ensuring that gradients are never zero during training.

The ReLU function and the Leaky ReLU function are nearly identical as seen in Fig. 6. During optimization, the Leaky ReLU foregoes hard-zero sparsity in exchange for a potentially more robust gradient. Alpha is a constant value (float >= 0).

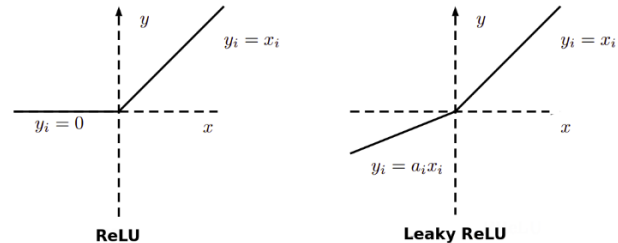


Fig. 6. ReLU vs Leaky ReLU [45].

Unlike the ReLU function, the Leaky ReLU has a non-zero gradient across its entire domain. The Leaky ReLU activation function is only available in the form of layers, not activations.

## IV. EXPERIMENTAL PROCEDURE

As shown in Fig. 7, the experiment design started with data collection and ended with a model evaluation.

### A. Data Collection

A dataset of river level and rainfall at Rantau Panjang, Pasir Mas in Kelantan from 2013 until 2017 was used, as shown in Fig. 8 and Fig. 9. The data of these rivers were recorded every year with flood occurrence [46]. The data were provided by the Department of Irrigation and Drainage (DID) Malaysia, and the features variable is shown in Table II. The river at Pasir Mas station collected river level (m) and rainfall (mm) daily. This dataset contained one measurement per day. The dependent variable was observed as a single daily value; hence, these cloud values were summed, accomplished by taking the average between 00:00 and 24:00 as each day's value. Because there were fewer and more irregular observations, the cloud base was summarised by averaging an entire day.

The dataset was imported into pandas by using the read csv() function and saved in the Data Frame named "df". Because the dataset was in tabular form, it was automatically converted into a Data Frame when working with tabular data in Pandas. In Python, a Data Frame is a two-dimensional, mutable data structure. It is made up of rows and columns, much like an excel sheet.

### B. Data Cleaning

Data cleaning or filtering's main function was to correct (or remove) and detect inaccurate data in the dataset. The task involves identifying inaccurate, incomplete, incorrect or irrelevant parts of the data and then deleting the noise data, and replacing or modifying them [48].



**Data Collection**

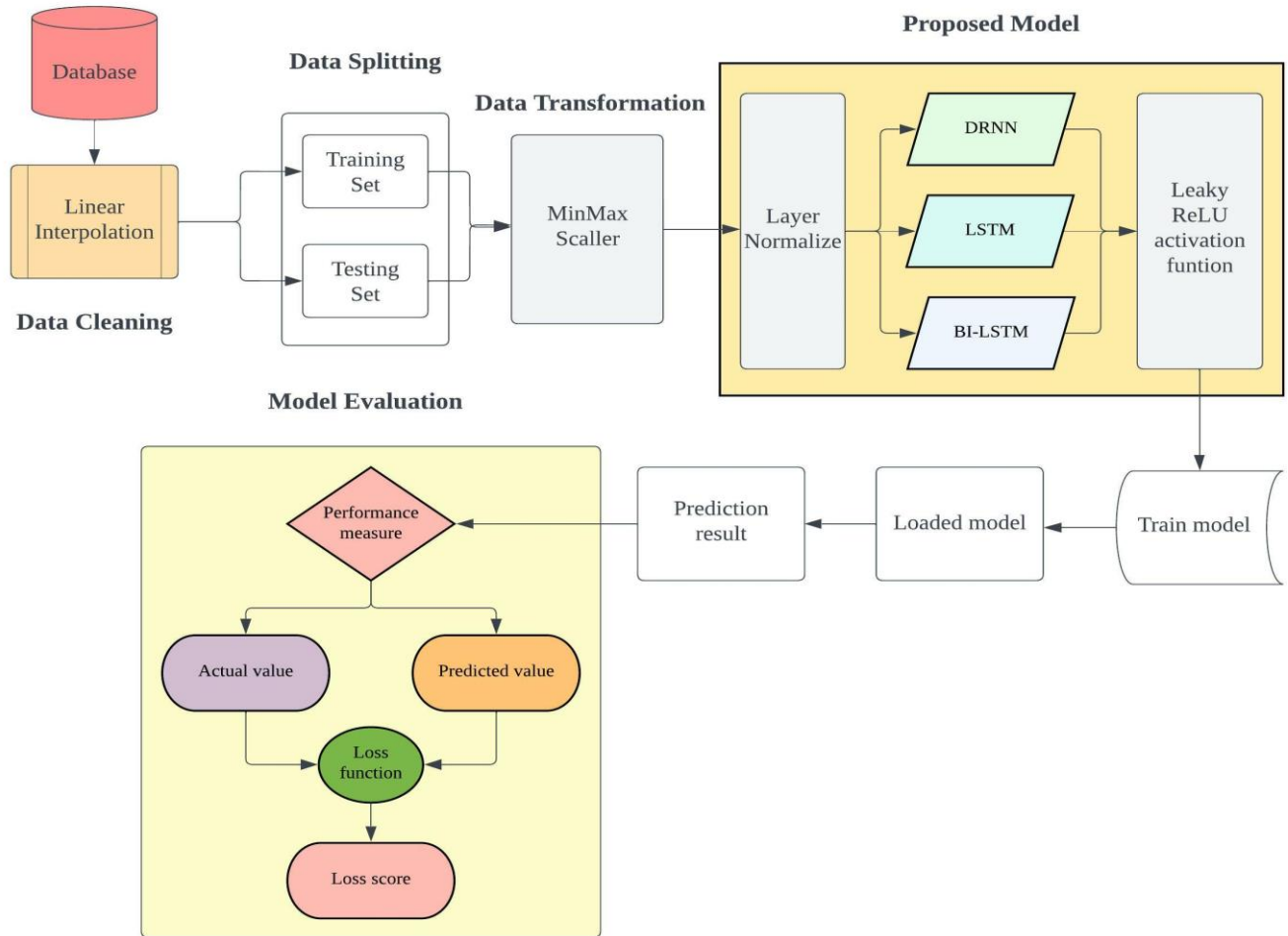


Fig. 7. Experiment Design.

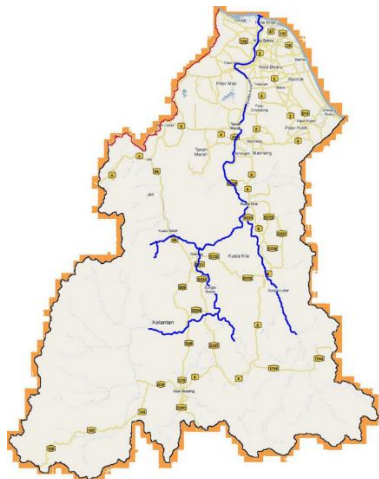


Fig. 8. Kelantan River Map [47].

AVERAGE AMOUNT RAINFALL PASIR MAS, KELANTAN  
IN 2014, 2106 AND 2018

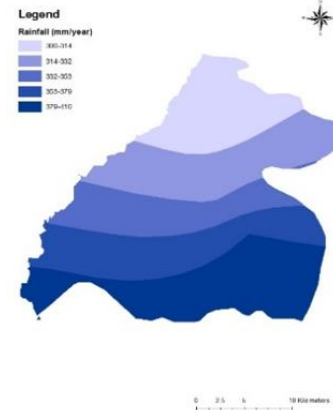


Fig. 9. Average Rainfall in Pasir Mas [46].



TABLE II. SUMMARY OF VARIABLE DATA COLLECTION

Variable	Unit	Description
Date	Date/ dd/mm/yyyy	Describe the current date
River Stage	Meter/m	Describe the river level during the day
Rainfall	Millimetres/mm	Describe the rain during the day

Data cleaning is the most crucial task because having incorrect or poor-quality data can harm processes and analysis. Clean data boosts overall productivity and allows for the utilization of the highest quality data when making predictions.

First, the river of Pasir Mas dataset needed to be checked to determine whether the dataset had missing or null values. After reviewing the dataset, a missing or null value in this situation needed to be dealt with through two options: a) removed a row with a missing value; b) replaced the missing value by using one of the most fundamental methods namely linear interpolation method. The formula is as follows:

$$y = y_a + (y_b - y_a) \frac{x - x_a}{x_b - x_a} \text{ at the point } (x, y)$$

$$\frac{y - y_a}{y_b - y_a} = \frac{x - x_a}{x_b - x_a} \quad (11)$$

$$\frac{y - y_a}{x - x_a} = \frac{y_b - y_a}{x_b - x_a}$$

The equation of [49] can be referred to as (11), the new line slope between  $(x_a, y_a)$  and  $(x, y)$  is the same as the slope of the line between  $(x_a, y_b)$  and  $(x_b, y_b)$ . The advantage of linear interpolation is that it is easy and fast to be applied, but its accuracy is doubtful.

The river of Rantau Panjang Pasir Mas station had ten missing values in the river level variable because of the existence of the same reading of recording on certain days. Handling missing values was done by the python pandas.

The interpolate () function fills NA values or missing values in the series or data frame. Rather than hard-coding the missing value, various interpolation, and convenient techniques could be used to fill the missing value.

### C. Dataset Splitting

The dataset needed to be divided into training and testing sets to avoid any phenomenon such as overfitting. Besides, the size of the datasets as well as the train/test split ratios can significantly impact the model output, thus, affecting classification performance [50]. For example, if there are patterns in the training and testing set that do not exist in real-world data, the model performs poorly even though it cannot be seen in the performance evaluation. Dataset splitting is a practice considered indispensable and highly necessary to eliminate or reduce bias in training data for prediction models [51].

Based on the Rantau Panjang river dataset, 80% were training data, and the remaining 20% were testing data that would be optimally splitting the training and testing the dataset.

### D. Data Transformation

Normalization is a scaling, mapping technique, or pre-processing stage. For prediction or forecasting purposes, it can be useful when it distinguishes a new range from an existing one.

Normalization is a transformation process that utilises a standard scale to produce numerically and comparably input data. After collecting input data, the data should perform some pre-processing to make it worthwhile for decision modelling [52]. As previously stated, this pre-processing should take three critical factors into account: 1) removed missing values from the data; 2) converted all non-numeric data to numerical data to allow for normalization (standardisation); 3) Determined how to select a suitable normalization technique to ensure a standard scale, appropriate modelling representation (benefit or cost criteria), and aggregation comparability to obtain alternative ratings.

After the data cleaning process, river stage and rainfall data underwent a min-max scaler to get normalized data. Min-max scaler scaled the data between the minimum and maximum value of the data that ended up ranging between 0 and 1. Another function for normalizing the data was to speed up the learning time and performance of the model. The data were scaled down to a range between [0, 1] or [-1, 1]. The method's equation of [14] can be referred to as (12):

$$a_{\text{norm}} = \frac{(\text{high} - \text{low}) * (a - \text{minA})}{\text{maxA} - \text{minA}} \quad (12)$$

min A is the smallest value, and max A is the largest value of attribute A.

### E. Model Evaluation

1) MAE stands for Mean Absolute Error, which is

$$\frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (13)$$

Outliers are given less weight in this method, which is not sensitive to outliers. The equation of [53] can be referred to as (13).

2) MAPE stands for Mean Absolute Percentage Error, which is

$$\frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (14)$$

MAE is similar, but true observation is used to normalise it. The disadvantage is that this metric becomes problematic when true observation is zero. The equation of [53] can be referred to as (14).

3) MSE stands for Mean Squared Error, which is

$$\frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 \quad (15)$$

MSE is a combination of variance of the prediction and measurement of bias, i.e.,  $MSE = \text{Bias}^2 + \text{variance}$ . The equation of [53] can be referred to as (15).

4) RMSE stands for Root Mean Squared Error, which is

$$\sqrt{\frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2} \quad (16)$$

It measures the standard deviation of residuals. The equation of [53] can be referred to as (16).

5) R2 stands for coefficient of determination, which is

$$1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (17)$$

Representing the coefficient indicates how well the values fit in comparison to the original values. The equation of [54] can be referred to as (17).

### V. RESULTS AND DISCUSSIONS

The results between the proposed models and original models that share the same hyperparameter setting were compared as shown in Table III.

In the present study, the proposed models were evaluated and compared with the original RNN model proposed by Hochreiter and Schmidhuber [55], the LSTM model proposed by Rumelhart & McClelland [56] and the BI-LSTM model proposed by Graves & Schmidhuber [57] to get a better understanding of whether the proposed models could produce better results [58][59][60]. The proposed models had an extra layer called layer normalization and one activation function is known as Leaky ReLU. In contrast, the original models had a standard layer and used a sigmoid as its activation function. Each deep learning model needs to be compared to the proposed and original models to ascertain which deep learning models perform better [61][62].

TABLE III. COMPARISON RESULT

Models	MSE	MAE	RSME	MAPE (%)	R2	Training Time
DRNN + LN + Leaky ReLU	0.107	0.233	0.327	4.293	0.946	1.08 minute
DRNN	0.118	0.242	0.343	4.344	0.94	0.38 second
LSTM + LN + Leaky ReLU	0.125	0.231	0.353	4.117	0.936	1.33 minute
LSTM	0.122	0.236	0.349	4.296	0.931	1.32 minute
BI-LSTM + LN + Leaky ReLU	0.131	0.258	0.362	4.751	0.933	3.33 minute
BI-LSTM	0.123	0.243	0.351	4.432	0.937	2.39 minute

Table III shows that the proposed models produced a better accuracy result: for example, the Deep Recurrent Neural Network (DRNN) + LN + Leaky ReLU model produced the lowest MSE among other models whereas the BI-LSTM + LN + Leaky ReLU model produced the highest MSE. For MAE,

the LSTM + LN + Leaky ReLU model produced the lowest error and the BI-LSTM + LN + Leaky ReLU model produced the highest error among other models. For RSME, the DRNN + LN + Leaky ReLU model produced the lowest error whereas the BI-LSTM + LN + Leaky ReLU model produced the highest error. For MAPE, the interpretation was in percentage mode and the lower is better, where the LSTM + LN + Leaky ReLU model produced the lowest MAPE, and the BI-LSTM + LN + Leaky ReLU model produced the highest MAPE. For R2, the higher the interpretation is better, in which the DRNN + LN + Leaky ReLU model produced the highest R2 whereas the LSTM model produced the lowest R2.

Comparing the proposed models with the original models, the DRNN + LN + Leaky ReLU model produced a low error in terms of MSE, MAE, RSME, MAPE and R2 which indicated that the proposed models were good in making a prediction, but the drawback was that it took longer time to train the data. For the LSTM model, the LSTM + LN + Leaky ReLU model produced the lowest error in terms of MAE, MAPE and R2 which was slightly better than the LSTM model. For the BI-LSTM model, the BI-LSTM produced better results compared to the BI-LSTM + LN + Leaky ReLU model in terms of MSE, MAE, RSME, MAPE, R2 and training time. The proposed model could not work well with the BI-LSTM model.

Additionally, the BI-LSTM model required more training time compared to the other models. In addition to performance evaluation, training time must also be taken into account. For instance, it is clear that the proposed models needed more training time than the original model because it had an additional layer called the normalization layer.

In literature, the LSTM model is regarded as the best performer compared to other models developed with dependent variables. However, in this case, the DRNN model performance is better compared to other models. This result can be seen in Table III, in which the DRNN + LN + Leaky ReLU model outperform other models in terms of the MSE, RSME, R2 and training time. The LSTM model shows the second best among other deep learning models with the lowest MAE and MAPE with the LSTM + LN + Leaky ReLU model. The BI-LSTM model shows the lowest accuracy among other models, one thing needs to be highlighted the result has also shown that the accuracy among the models is about the same, the differences are just slight, and there is a considerable gap between them. The DRNN model is the first place in accuracy, followed by the LSTM model and the lowest is the BI-LSTM model. Because the LSTM framework uses backpropagation and a gate to train the model, it takes more time to train than the DRNN model, which uses sequential while training the data and has no backpropagation and gate in the architecture. The DRNN model ranks lowest in terms of training time, while the LSTM model comes in second. Contrarily, the BI-LSTM model necessitates that the training data move in both past and forward directions to train the data, which is why the BI-LSTM model takes longer to train the data.

Based on the result in Table III, only the proposed models with layer normalization and Leaky ReLU are deemed suitable for flood prediction with minimal missing value in the data usage which results in the lowest minimum error and good

accuracy. The missing value in the data can be filled by using the linear interpolation method to get the best possible clean data. The authorities may use these models as an alternative to anticipate flooding and make enough preparations prior to its occurrence.

## VI. CONCLUSION

In conclusion, the DRNN model performs relatively well compared to the LSTM and BI-LSTM models with the used dataset. From the literature, the LSTM architecture needs requirements for backpropagation and a gate to train the model. Therefore, the LSTM model is marginally more complicated than the DRNN model. Meanwhile, the BI-LSTM model performs with somewhat lower accuracy but is still able to deliver a good outcome. Additionally, the BI-LSTM model requires the training data to move backwards and forward in both directions which increases the training time needed. Even though the performance of proposed models performs well, there are still many improvements that can be made using deep learning approaches.

## VII. FUTURE WORK

For future work, it is suggested that additional experiments be conducted by combining a statistic model with the pre-processing models to ascertain how the combined model performs. Currently, these models produce a good accuracy for one day ahead but for future work these models need to be tuned to produce a good accuracy for multi-days ahead.

## ACKNOWLEDGMENT

The authors are grateful to Universiti Teknikal Malaysia Melaka for the financial support through the university's short-term grant PJP/2020/FTMK/PP/S01802.

## REFERENCES

- [1] T. Ashizawa, N. Sudo, and H. Yamamoto, "How Do Floods Affect the Economy? An Empirical Analysis using Japanese Flood Data," 2022.
- [2] T. Tiggeloven et al., "Global-scale benefit-cost analysis of coastal flood adaptation to different flood risk drivers using structural measures," *Nat. Hazards Earth Syst. Sci.*, vol. 20, no. 4, pp. 1025–1044, 2020, doi: 10.5194/nhess-20-1025-2020.
- [3] M. S. M. Shaari, M. Z. Abd Karim, and B. Hasan-Basri, "Does flood disaster lessen GDP growth? Evidence from Malaysia's manufacturing and agricultural sectors," *Malaysian J. Econ. Stud.*, vol. 54, no. 1, pp. 61–81, 2017, doi: 10.22452/mjes.vol54no1.4.
- [4] K. A. Oladapo, S. A. Idowu, Y. . Adekunle, and F. . Ayankoya, "Categorization of Conditioning Variables for Pluvial Flood Risk Assessment," *Int. J. Sci. Eng. Res.*, vol. 11, no. 8, pp. 355–368, 2020.
- [5] S. F. Zakaria, R. M. Zin, I. Mohamad, S. Balubaid, S. H. Mydin, and E. M. R. Mdr, "The development of flood map in Malaysia," *AIP Conf. Proc.*, vol. 1903, no. November, 2017, doi: 10.1063/1.5011632.
- [6] F. Nurashikin Sungip et al., "The Impact of Monsoon Flood Phenomenon on Tourism Sector in Kelantan, Malaysia: A Review," *Int. J. Eng. Technol.*, vol. 7, no. 4.34, p. 37, 2018, doi: 10.14419/ijet.v7i4.34.23577.
- [7] N. S. Romali, Z. Yusop, M. Sulaiman, and Z. Ismail, "Flood risk assessment: A review of flood damage estimation model for Malaysia," *J. Teknol.*, vol. 80, no. 3, pp. 145–153, 2018, doi: 10.11113/jt.v80.11189.
- [8] M. Zerara, "Machine Learning and Statistical Methods for Time Series Forecasting : a Case Machine Learning and Statistical Methods for Time Series Forecasting : a Case Study with Water Demand," no. March, 2021, doi: 10.36227/techrxiv.15019698.v1.
- [9] A. Pant and R. S. Rajput, "Time Series Analysis of Gold Price Using R," no. January, 2018.
- [10] M. Murat, I. Malinowska, M. Gos, and J. Krzyszczyk, "Forecasting daily meteorological time series using ARIMA and regression models," *Int. Agrophysics*, vol. 32, no. 2, pp. 253–264, 2018, doi: 10.1515/intag-2017-0007.
- [11] R. Yasrab, J. Zhang, P. Smyth, and M. P. Pound, "Predicting plant growth from time-series data using deep learning," *Remote Sens.*, vol. 13, no. 3, pp. 1–17, 2021, doi: 10.3390/rs13030331.
- [12] S. Joshi, "Time Series Analysis and Forecasting of the US Housing Starts using Econometric and Machine Learning Model," no. May, 2019.
- [13] A. T. Jebb and L. Tay, *Introduction to Time Series Analysis for Organizational Research: Methods for Longitudinal Analyses*, vol. 20, no. 1, 2017.
- [14] S. Bhanja and A. Das, "Deep Neural Network for Multivariate Time-Series Forecasting," *Adv. Intell. Syst. Comput.*, vol. 1255, no. April, pp. 267–277, 2021, doi: 10.1007/978-981-15-7834-2\_25.
- [15] W. Wu, R. Emerton, Q. Duan, A. W. Wood, F. Wetterhall, and D. E. Robertson, "Ensemble flood forecasting: Current status and future opportunities," *WIREs Water*, vol. 7, no. 3, 2020, doi: 10.1002/wat2.1432.
- [16] Z. Shen, Y. Zhang, J. Lu, J. Xu, and G. Xiao, "A novel time series forecasting model with deep learning," *Neurocomputing*, vol. 396, no. 0925–2313, pp. 302–313, 2020, doi: <https://doi.org/10.1016/j.neucom.2018.12.084>.
- [17] C. L. Jun, Z. S. Mohamed, A. L. S. Peik, S. F. M. Razali, and S. Sharil, "Flood forecasting model using empirical method for a small catchment area," *J. Eng. Sci. Technol.*, vol. 11, no. 5, pp. 666–672, 2016.
- [18] A. Faruq, H. P. Arsa, S. F. M. Hussein, C. M. C. Razali, A. Marto, and S. S. Abdullah, "Deep Learning-Based Forecast and Warning of Floods in Klang River, Malaysia," *Ing. des Syst. d'Information*, vol. 25, no. 3, pp. 365–370, 2020, doi: 10.18280/isi.250311.
- [19] S. A. Asmai, Z. Z. Abidin, H. Basiron, and S. Ahmad, "An intelligent crisis-mapping framework for flood prediction," *Int. J. Recent Technol. Eng.*, vol. 8, no. 2 Special Issue 8, pp. 1304–1310, 2019, doi: 10.35940/ijrte.B1058.0882S819.
- [20] S. Siami-Namini and A. S. Namin, "Forecasting Economics and Financial Time Series: ARIMA vs. LSTM," pp. 1–19, 2018.
- [21] J. K. Jaiswal and R. Das, "Artificial neural network algorithms based nonlinear data analysis for forecasting in the finance sector," *Int. J. Eng. Technol.*, vol. 7, no. 4, pp. 169–176, 2018, doi: 10.14419/ijet.v7i4.10.20829.
- [22] A. Šiljić Tomić, D. Antanasijević, M. Ristić, A. Perić-Grujić, and V. Pocajt, "A linear and non-linear polynomial neural network modeling of dissolved oxygen content in surface water: Inter- and extrapolation performance with inputs," *Sci. Total Environ.*, vol. 610–611, pp. 1038–1046, 2018, doi: 10.1016/j.scitotenv.2017.08.192.
- [23] Y. F. Zhang, P. Fitch, and P. J. Thorburn, "Predicting the trend of dissolved oxygen based on the PCA-RNN model," *Water (Switzerland)*, vol. 12, no. 2, 2020, doi: 10.3390/w12020585.
- [24] M. Coto-Jiménez, "Experimental study on transfer learning in denoising autoencoders for speech enhancement," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 12088 LNCS, pp. 307–317, 2020, doi: 10.1007/978-3-030-49076-8\_29.
- [25] A. Z. Dhunny, R. H. Seebocus, Z. Allam, M. Y. Chuttur, M. Eltahan, and H. Mehta, "Flood Prediction using Artificial Neural Networks: Empirical Evidence from Mauritius as a Case Study," *Knowl. Eng. Data Sci.*, vol. 3, no. 1, pp. 1–10, 2020, doi: 10.17977/um018v3i12020p1-10.
- [26] V. Kumar and M. L., "Deep Learning as a Frontier of Machine Learning: A Review," *Int. J. Comput. Appl.*, vol. 182, no. 1, pp. 22–30, 2018, doi: 10.5120/ijca2018917433.
- [27] A. Mathew, P. Amudha, and S. Sivakumari, "Deep learning techniques: an overview," *Adv. Intell. Syst. Comput.*, vol. 1141, no. January, pp. 599–608, 2021, doi: 10.1007/978-981-15-3383-9\_54.
- [28] R. Vargas, Rocio, Mosavi, Amir, & Ruiz, "Deep Learning : a Review Deep Learning : a Review," *Adv. Intell. Syst. Comput.*, no. July, 2017.
- [29] J. Fu, J. Chu, P. Guo, and Z. Chen, "Condition Monitoring of Wind Turbine Gearbox Bearing Based on Deep Learning Model," *IEEE Access*, vol. 7, no. April, pp. 57078–57087, 2019, doi: 10.1109/ACCESS.2019.2912621.

- [30] L. Hu, J. Zhang, Y. Xiang, and W. Wang, "Neural Networks-Based Aerodynamic Data Modeling: A Comprehensive Review," *IEEE Access*, vol. 8, pp. 90805–90823, 2020, doi: 10.1109/ACCESS.2020.2993562.
- [31] G. Chen, "A Gentle Tutorial of Recurrent Neural Network with Error Backpropagation," pp. 1–10, 2016.
- [32] H. H. Hasan, S. F. M. Razali, A. Z. I. A. Zaki, and F.-3. pdfu. M. Hamzah, "Integrated hydrological-hydraulic model for flood simulation in tropical urban catchment," *Sustain.*, vol. 11, no. 23, 2019, doi: 10.3390/su11236700.
- [33] B. Biswal, "Hydrological modelling: scale issues and philosophical approaches," no. May, pp. 36–50, 2019.
- [34] E. Brown et al., "Methods and tools to support real time risk-based flood forecasting - A UK pilot application," *E3S Web Conf.*, vol. 7, no. October, 2016, doi: 10.1051/e3sconf/20160718019.
- [35] M. Bayat and O. Tavakkoli, "Application of machine learning in flood forecasting," *Futur. Technol.*, vol. 1, no. 1, pp. 1–6, 2020, doi: 10.55670/fpll.futech.1.1.1.
- [36] X. H. Le, H. V. Ho, G. Lee, and S. Jung, "Application of Long Short-Term Memory (LSTM) neural network for flood forecasting," *Water (Switzerland)*, vol. 11, no. 7, 2019, doi: 10.3390/w11071387.
- [37] P. Goymann, D. Herrling, and A. Rausch, "Flood Prediction through Artificial Neural Networks A case study in Goslar , Lower Saxony," no. c, pp. 56–62, 2019.
- [38] V. Gude, S. Corns, and S. Long, "Flood Prediction and Uncertainty Estimation Using Deep Learning," *Water (Switzerland)*, vol. 12, no. 3, 2020, doi: 10.3390/w12030884.
- [39] T. Song, W. Ding, J. Wu, H. Liu, H. Zhou, and J. Chu, "Flash flood forecasting based on long short-term memory networks," *Water (Switzerland)*, vol. 12, no. 1, 2020, doi: 10.3390/w12010109.
- [40] S. Bai, J. Z. Kolter, and V. Koltun, "An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling," 2018.
- [41] P. Lara-Benítez, M. Carranza-García, J. M. Luna-Romera, and J. C. Riquelme, "Temporal convolutional networks applied to energy-related time series forecasting," *Appl. Sci.*, vol. 10, no. 7, 2020, doi: 10.3390/app10072322.
- [42] X. Bi and L. Wang, "Performing Weakly Supervised Retail Instance Segmentation via Region Normalization," *IEEE Access*, vol. 9, pp. 67761–67775, 2021, doi: 10.1109/ACCESS.2021.3077031.
- [43] J. Xu, X. Sun, Z. Zhang, G. Zhao, and J. Lin, "Understanding and improving layer normalization," *Adv. Neural Inf. Process. Syst.*, vol. 32, no. November, 2019.
- [44] G. Renith and A. Senthilselvi, "Accuracy improvement in diabetic retinopathy detection using dlia," *J. Adv. Res. Dyn. Control Syst.*, vol. 12, no. 4, pp. 133–149, 2020, doi: 10.5373/JARDCS/V12I4/20201426.
- [45] S. Nasiri, J. Helsper, M. Jung, and M. Fathi, "DePicT Melanoma Deep-CLASS: A deep convolutional neural networks approach to classify skin lesion images," *BMC Bioinformatics*, vol. 21, no. Suppl 2, pp. 1–14, 2020, doi: 10.1186/s12859-020-3351-y.
- [46] N. A. Mohamd Hanan et al., "A GIS-Based Flood Vulnerability Assessment in Pasir Mas, Kelantan," *IOP Conf. Ser. Earth Environ. Sci.*, vol. 549, no. 1, 2020, doi: 10.1088/1755-1315/549/1/012004.
- [47] N. Ahmad Zamree, N. A. Said, and S. Sibly, "Hospital Disaster Preparedness: A Model for Hospital Disaster Preparedness Based on 2014 Flood in Kelantan," *Educ. Med. J.*, vol. 10, no. 4, pp. 69–80, 2019, doi: 10.21315/eimj2018.10.4.7.
- [48] F. Ridzuan and W. M. N. Wan Zainon, "A review on data cleansing methods for big data," *Procedia Comput. Sci.*, vol. 161, pp. 731–738, 2019, doi: 10.1016/j.procs.2019.11.177.
- [49] P. Xu and Y. Jia, "SNR improvement based on piecewise linear interpolation," *J. Electr. Eng.*, vol. 72, no. 5, pp. 348–351, 2021, doi: 10.2478/jee-2021-0049.
- [50] A. Rác, D. Bajusz, and K. Héberger, "Effect of dataset size and train/test split ratios in qsar/qspr multiclass classification," *Molecules*, vol. 26, no. 4, pp. 1–16, 2021, doi: 10.3390/molecules26041111.
- [51] I. Muraina, "IDEAL DATASET SPLITTING RATIOS IN MACHINE LEARNING ALGORITHMS :," no. February, 2022.
- [52] B. Etzkorn, "Data Normalization and Standardization," pp. 1–3, 2018.
- [53] L. Chen, X. Yang, C. Sun, Y. Wang, D. Xu, and C. Zhou, "Feed intake prediction model for group fish using the MEA-BP neural network in intensive aquaculture," *Inf. Process. Agric.*, vol. 7, no. 2, pp. 261–271, 2020, doi: 10.1016/j.inpa.2019.09.001.
- [54] B. Panay, N. Baloian, J. Pino, S. Peñafiel, H. Sanson, and N. Bersano, "Predicting Health Care Costs Using Evidence Regression," p. 74, 2019, doi: 10.3390/proceedings2019031074.
- [55] W. Feng, N. Guan, Y. Li, X. Zhang, and Z. Luo, "Audio visual speech recognition with multimodal recurrent neural networks," *Proc. Int. Jt. Conf. Neural Networks*, vol. 2017-May, no. May, pp. 681–688, 2017, doi: 10.1109/IJCNN.2017.7965918.
- [56] Z. Yu and G. Liu, "Sliced recurrent neural networks," *COLING 2018 - 27th Int. Conf. Comput. Linguist. Proc.*, pp. 2953–2964, 2018.
- [57] T. Zhu, J. Shen, and F. Sun, "Long Short-Term Memory-based simulation study of river happiness evaluation – A case study of Jiangsu section of Huaihe River Basin in China," *Heliyon*, vol. 8, no. 9, p. e10550, 2022, doi: 10.1016/j.heliyon.2022.e10550.
- [58] Ammar Ashraf Narul Akhla, Thong Chee Ling, Abdul Samad Shibghatullah, Chit Su Mon, Aswani Kumar Cherukuri, Chaw Lee Yen and Lee Chiu Yi, 2022, "Impact of Real-Time Information for Travellers: A Systematic Review", *Journal of Information & Knowledge Management*, vol. 21, no.4, pp. 2250065-1-2250065-21, 2022.
- [59] Susanto, I.C., Subaramaniam, K., Shibghatullah, A.S.B., "Gesturenomy: Touchless Restaurant Menu Using Hand Gesture Recognition", *Proceedings of International Conference on Artificial Life and Robotics*, pp.229-236, 2022.
- [60] Meng, W.Y., Shibghatullah, A.S.B., Subaramaniam, K., "Smart Tourism Guide Application Using Location-Based Services-Go.Travel", *Proceedings of International Conference on Artificial Life and Robotics*, pp. 219-228, 2022.
- [61] Shibghatullah, A.S., Jalil, A., Wahab, M.H.A., Soon, J.N.P., Subaramaniam, K., Eldabi, T., "Vehicle Tracking Application Based on Real Time Traffic", *International Journal of Electrical and Electronic Engineering and Telecommunications*, vol. 11, no. 1, pp. 67-73, 2022
- [62] Haini, M.S.B., Mon, C.S., Shibghatullah, A.S.B., Jalil, A.B., Subaramaniam, K.A.P., Hussin, A.A.A., "An investigation into requirement of mobile app for apartment residents", *International Journal on Advanced Science, Engineering and Information Technology*, vol. 9, no. 6, pp. 1841-1848, 2019.

# Recognition Method of Dim and Small Targets in SAR Images based on Machine Vision

Qin Dong

Yancheng Institute of Technology  
School of Information Engineering, Yancheng 224000, China

**Abstract**—Aiming at the problems of long recognition time and low recognition accuracy of traditional SAR image dim target recognition methods, a method of SAR image dim target recognition based on machine vision was proposed. SAR images are collected and preprocessed by machine vision, and the image information is processed by PCA dimension reduction considering the linear characteristics of the data to extract image features. Then, the SAR image target feature key frame frequency band is divided by the segmentation results, and the recognition model is established based on the image trajectory tracking and target analysis. The proposed algorithm is applied and analyzed. The simulation results show that the proposed algorithm has good recognition rate, average recognition rate and false detection rate are 99% and 0.9%, and can effectively ensure the data processing performance.

**Keywords**—Machine vision; SAR image; Weak target; PCA linear dimensionality reduction method; key frame frequency band

## I. INTRODUCTION

Synthetic Aperture Radar (SAR) is an active microwave imaging radar. SAR has high range and azimuth resolution, and can obtain two-dimensional high-resolution images. Since the electromagnetic waves emitted by SAR have a longer wavelength and can penetrate clouds, fog, rain, smoke, haze, etc., SAR has a certain penetration ability and can work all day and all day long. Based on the above advantages, SAR is widely used in the military field, mainly used for: military intelligence reconnaissance and tracking, military terrain mapping, navigation and guidance, damage evaluation, etc. [1]. In addition, with the further maturity of SAR technology, SAR is also widely used in many civilian fields, such as resource monitoring, environmental monitoring, forest vegetation cover detection, disaster monitoring, and ecology, hydrology, archaeology, and deep space exploration. With the continuous development of SAR technology, the resolution of radars is getting higher and higher, the quality of SAR images is getting better and better, and new SAR images are constantly being produced [2]. What does not match it is that the interpretation of SAR images has developed slowly. The initial manual interpretation is completely performed by personnel based on the shape, size, shadow, position, and hue of the target in the image to extract the information in the SAR image. This method is not only extensive, the knowledge and the deep background of SAR, and the efficiency is very low, it is greatly affected by the subjective experience of the observer, and it is far from satisfying the current situation of massive SAR images that need to be analyzed. Target recognition in SAR images does not require manual intervention, and relies on computers to run

related algorithms to automatically classify SAR image targets. The general target recognition step is: first extract the characteristics of the image target, and then recognize it through the corresponding recognition method [3].

Dim and small target recognition in SAR image is one of the core technologies in the automatic target detection system. When the distance is far, the imaging area of these targets on the focal plane is very small, generally not exceeding the size of the detector pixel. The target appears as dots in the image, and the signal-to-noise ratio is extremely low. The target is submerged by noise, which brings great difficulties to target detection. For more than ten years, the identification of small and dim targets in SAR images has been a research hotspot in the field of optics and infrared images [4]. The International Optical Engineering Society organizes an annual “Signal and Data Processing of Small Targets” conference to exchange new technologies for the recognition of small and small targets in SAR images. The International Optical Engineering Society defines the weak target as: the image size is less than 80 pixels, that is, less than 0.15% of 256×256. The research on the recognition of small and dim targets in SAR images originated from long-distance search and surveillance. For example, use wide-field telescopes to search for or track meteors, satellites or other moving targets in the sky, and use airborne or ground infrared (TV) search and tracking systems to search for long-distance targets. It uses image processing algorithms to automatically recognize targets in a cluttered background and strong noise environment. The performance of the algorithm is critical to the range and intelligence of the automatic target recognition system. Recognition of small and dim targets in SAR images is a difficult subject with important strategic application value [5].

At the same time, for the research of SAR target recognition algorithm, most scholars try to improve the algorithm from different angles to improve its target recognition accuracy, thereby reducing background clutter and noise interference factors in image recognition. Literature [6] proposed that the one-dimensional feature extraction of principal component analysis is used as the input data of the encoder, and then the SAR target image is used as the input of the neural network to realize the target recognition of two-dimensional data, and research has proved that the depth learning algorithm of fusion decision layer and feature layer has good adaptability and robustness. In reference [7], SAR image classification method based on target decomposition and support vector machine is used to decompose and combine features in polarization and establish polarization classification model. The experimental

results show that the improved polarization image classification method has better classification performance and application effectiveness. Literature [8] proposed a view tensor sparse representation model based on target recognition, and used JT-OMP algorithm to calculate the sparse representation error of SAR tensor image data and multi view SAR image after the construction of recognition dictionary. Subsequently, the effectiveness of the algorithm was verified in the target recognition database. In reference [9], considering the decline in accuracy of ship target image due to the motion of the target, it proposed to combine clustering algorithm with SAR image recognition, and proposed HCA ship target aggregation algorithm in airborne SAR image. The distance algorithm is used as the basis for the generation of a single image. The experimental results show that this method can have good application performance in simulation experiments. Reference [10] proposed a SAR target recognition method based on adaptive kernel dictionary learning, that is, extracting nonlinear feature information through data space mapping, information dynamic updating, and minimizing error reconstruction. Simulation results show that this method has good recognition performance. In reference [11], zero phase component analysis is used to achieve feature extraction, and sparse technology is used to optimize the convolutional neural network SAR target recognition algorithm. The results show that the algorithm has high target recognition ability and good noise robustness.

Literature [12] designed a multi-azimuth SAR image for target recognition convolutional neural network (Convolutional Neural Network, CNN), three SAR images of the same target are input into the network as a pseudo-color image, making full use of the characteristics of SAR image data acquisition are improved, and the flattening operation is replaced by a pooling layer, which reduces the number of network parameters. The experimental results show that even on a small-scale SAR data set, the convolutional network has the characteristics of high recognition accuracy. Targets of different models in the same category also have excellent recognition performance. Literature [13] proposed a two-dimensional principal component analysis (2DPCA) and L2 regularization constrained stochastic configuration network (SCN) for integrated learning SAR image target recognition method, 2DPCA not only It can effectively extract the feature information of the target and reduce the amount of data through sparse representation. The SCN regularization algorithm has fewer parameters and can effectively avoid the network overfitting problem and improve the recognition rate of the network. Although the above two methods realize the recognition of small and weak targets in SAR images, they take a long time to recognize and the recognition efficiency is low. Literature [14] proposes a synthetic aperture radar (SAR) image target recognition method based on random weighted fusion of single-level signal decision-making layers, and uses sparse representation classification (SRC) to implement the multi-level and multi-component single-level signal representation obtained from SAR image decomposition. For decision-making, the error vector is fused by a random weight matrix, which contains a large number of random weights. According to the fusion results, different types of error statistics can be obtained. The decision variables are defined to reflect the correlation of different types. Finally, according to the minimum error is used

to make category decision. Extensive experiments are carried out on the MSTAR data set and compared with many types of existing methods. The results show that the proposed method can effectively improve the overall performance of SAR target recognition. Literature [15] proposed a Capsule-based SAR image target recognition method, which uses multiple convolutional layers to achieve hierarchical processing, while using fewer convolution kernels, but the number of convolution kernels used in each layer gradually increases as the level deepens, which makes the extracted features more abstract. In the Primary Caps layer, the Capsule vector is composed of all the feature maps output by the last layer of the convolutional layer, so that the Capsule unit contains all the features of the target part or the whole to complete the complete instance of the target to achieve SAR image target recognition. However, the accuracy of the above two methods for small and weak target recognition in SAR images is low, resulting in poor recognition effect.

When the above methods are used for SAR target image recognition, they only focus on improving the accuracy of image recognition, but are difficult to analyze the characteristics of weak targets in SAR images. There are many factors involved in target weakening, including sensor itself, target scattering characteristics, background environmental factors and the amount of information about the target's environment. Based on the research limitations of previous scholars, the research proposes to use machine vision to recognize small and weak targets in SAR images. The image recognition is achieved through visual image acquisition - data preprocessing - feature extraction under linear dimension reduction - key frame division of target features. This not only ensures the effective detection of weak targets, but also considers the linear nature of data features, which improves the accuracy of algorithm recognition to a certain extent. At the same time, experimental simulation is used to verify the effectiveness of the algorithm, in order to provide a new idea for weak target image recognition.

## II. METHOD FOR RECOGNIZING DIM AND SMALL TARGETS IN SAR IMAGE

### A. SAR Image Acquisition based on Machine Vision

High-quality SAR images can better reflects the characteristics of the recognition target. In order to facilitate subsequent recognition operations and reduce the computational burden of the recognition algorithm, the use of machine vision imaging technology to collect SAR images enables the recognition algorithm to have relatively high recognition efficiency and recognition accuracy [16].

In the SAR image acquisition, the camera CCD is mainly used to convert the SAR signal into an orderly SAR signal and collect the target information to be identified. Considering the size of the target to be measured, as well as the imaging area, depth of field, working distance and other project requirements, the XF-5MDT05X65 telecentric lens is selected to cooperate with the camera to collect the image of the target to be recognized. The use of cameras to capture images has very strict requirements on the light source. When collecting SAR images, according to specific needs, choose a lighting plan to obtain the best lighting effect and obtain high-quality images. Considering



that the recognition target is a small and weak target in SAR image, the forward illumination method is selected, and the LTLNC100-W linear light source is used to light it [17]. For detection targets with complex structures and relatively large targets, it is easy to have uneven features during the acquisition process. In order to ensure that the characteristics of different regions in the collected SAR images are uniform, multiple light sources are used to illuminate different regions to improve the SAR image. The gray value is uniform level. Multiple light sources mainly illuminate the target in the form of superposition of light fields, as shown in Fig. 1.

The top light source shown in the upper part of Fig. 1 is mainly aimed at the center area of the target, and the imaging of the edge area is achieved through the bottom light source. When two light sources are used at the same time, the image gray value can be relatively uniform and improved. The quality of SAR images lays the foundation for the subsequent identification of small and dim targets in SAR images [18].

### B. SAR Image Preprocessing

Hu's invariant moments can solve the problem of image lens distortion in the image acquisition process, and improve the authenticity of image geometric moments. Then by inverting the radial distortion model of the image, the pixel coordinates of the image are restored, and the mapping relationship between the image pixel coordinates is grayed out [19].

Before determining the geometric shape characteristics of the SAR image, it is necessary to clarify the  $\alpha + \beta$ -order geometric moments contained in the SAR image:

$$m_{\alpha\beta} = \sum_{x=1}^M \sum_{y=1}^N a^\alpha b^\beta f(a,b); \alpha, \beta = 0, 1, 2, \dots \quad (1)$$

According to the calculation result of the above formula, it is known that the SAR image is  $M \times N$ , and  $f(a,b)$  is described as the actual gray value of the image, where  $a$  and  $b$  respectively represent the coordinate axis of the pixel point [20].

Suppose that in the discrete state of rotation, the sixth-order moments of translation, scaling, and rotation of  $\alpha + \beta$ , which contains geometric features, are derived from the ideal result of no distortion of geometric moments in the above-mentioned pattern. However, compared with the ideal result, the real SAR image must have a certain degree of distortion. This part of the error will make the pixel coordinate  $(a,b)$  and the gray level  $f(a,b)$  in the image unable to accurately correspond, resulting in geometric deformation of the SAR image after conversion [21].

To this end, this paper corrects the  $\alpha + \beta$ -order geometric moment of the SAR image by reversing the resolution of the radial camera distortion model.

Generally, the radial distortion model is defined as:

$$\begin{cases} a = \bar{a} + (\bar{a} - a_0)(O_1 r^2 + O_2 r^4 + \dots) \\ b = \bar{b} + (\bar{b} - b_0)(O_1 r^2 + O_2 r^4 + \dots) \end{cases} \quad (2)$$

According to the calculation results of the above formula,  $(a,b)$  is described as the actual pixel coordinate value due to the influence of distortion,  $(\bar{a}, \bar{b})$  can describe the ideal pixel coordinate unit, and  $(a_0, b_0)$  is described as the global center pixel coordinate [22]. Use  $O_i$  to describe the distortion coefficient produced by the  $2i$ -level of the image, and set the distortion influence range to be related to the distance  $r$  from the target point to  $(a_0, b_0)$ , and then:

$$r^2 = (a - a_0)^2 + (b - b_0)^2 \quad (3)$$

It is concluded that the geometric deformation in the SAR image collected by machine vision can be described by the second-order radial distortion coefficient  $K_i$ , without considering the influence of higher-order distortion terms [23]. Therefore, the ideal pixel coordinate  $(\bar{a}, \bar{b})$  is inversely solved by the simplified distortion calculation model obtained:

$$\begin{cases} \bar{a} = \frac{a + a_0 O_1 r^2}{1 + O_1 r^2} \\ \bar{b} = \frac{b + b_0 O_1 r^2}{1 + O_1 r^2} \end{cases} \quad (4)$$

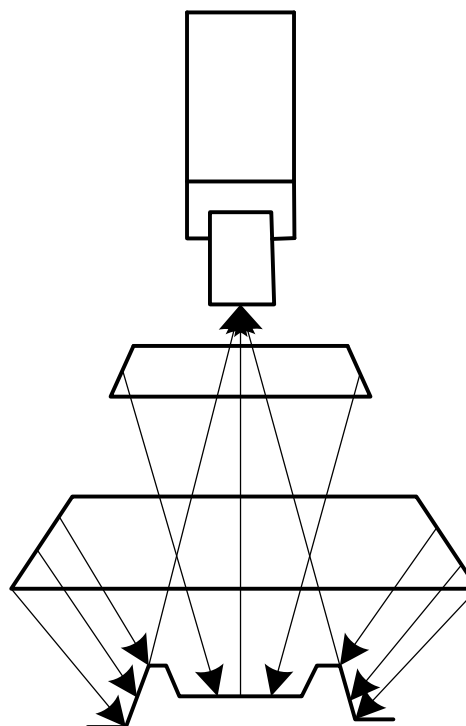


Fig. 1. SAR Image Imaging Scheme.

Bring the ideal pixel coordinates into equation (1) to obtain the corrected  $\alpha + \beta$  -order geometric moment of the SAR image:

$$m_{\alpha\beta} = \sum_{a=1}^M \sum_{b=1}^M \left( \frac{a + a_0 O_1 r^2}{1 + O_1 r^2} \right) \left( \frac{b + b_0 O_1 r^2}{1 + O_1 r^2} \right) \quad (5)$$

The combination of formula (5) and formula (3) can obtain the Hu stock moment with translation, zoom and rotation functions, which can increase the influence factor of radial distortion during the acquisition process, and reduce the deformation caused by the acquisition of SAR images by the hardware device. The problem of reduced recognition accuracy is discussed in [24].

### C. SAR Image Feature Extraction

Feature extraction is an important research problem in SAR image target recognition. Choosing appropriate features can improve target recognition rate and timeliness. The definition of feature extraction can be divided into a narrow sense and a broad sense. Feature extraction in a broad sense refers to a transformation that uses various mathematical transformation methods to improve the distribution of original features in the feature space without changing the internal structure and parameters. It can compress feature dimensions, remove redundant features, and reduce calculations. Effect: Use the feature space transformation method commonly used in pattern recognition for feature extraction [25]. PCA is a commonly used linear dimensionality reduction method in pattern recognition. PCA takes the maximum change direction of the sample in a multi-dimensional space (that is, the direction of maximum variance) as the criterion for judging whether the vector is a principal vector according to the position distribution of the sample in the space. Realize sample compression and feature extraction. Suppose the projection of vector  $x$  is  $y$ ,  $w$  is the projection matrix, and D can be represented by the inner product of  $x$  and  $w$ , namely:

$$y = [w, x] = \sum_{i=1}^n w_i x_i m_{\alpha\beta} = x^T x m_{\alpha\beta} \quad (6)$$

The purpose of PCA is to find an  $w$  that maximizes the value of variance  $E[y^2]$ ,  $E[y^2]$  can be expressed as:

$$E[y^2] = E[(w^T x)^2] = w^T E(xx^T)w = w^T C_x w \quad (7)$$

According to the theoretical knowledge of linear algebra, if the value of variance  $E[y^2]$  is the largest,  $E[y^2]$  can be expressed as:

$$E[y^2] = E[(w^T x)^2 \lambda_i] = w^T E(xx^T) \lambda_i = w^T C_x \lambda_i \quad (8)$$

$E[y^2]$ , which maximizes the value of  $w$  according to equation (8), is the eigenvector corresponding to the maximum eigenvalue of matrix  $C_x$ . for the component of eigenvalue  $\lambda_i$ , the variance of the principal component is also  $\lambda_i$ , which represents the dispersion degree of the sample in the direction

of the principal component. Data dimensionality reduction is realized by controlling the contribution  $n_i$  of principal component  $\lambda_i$  to the data, and  $n_i$  can be expressed by the following formula:

$$n_i = \frac{\lambda_i}{\lambda_1 + \lambda_2 + \dots + \lambda_p} \quad (9)$$

Generalized feature extraction methods basically have mature implementation algorithms, which have been widely used in SAR image target recognition. In a narrow sense, feature extraction is the process of extracting features that can reflect the essential attributes of SAR image targets [26].

### D. Key Frame Frequency Band Division of Target Feature in SAR Image

According to the above extracted SAR image features, the key frame frequency band of SAR image target features is divided. Firstly, the template matching method is used to construct the pixel feature point block matching structure model of SAR image target, as shown in Fig. 2.

In the block matching structure model shown in Fig. 2, the pixel frame is used to match the template of SAR image target. For the target image collected in the  $k$ -th subband, the key frame fusion method is used to construct the frequency band division model of SAR image target. According to the correlation between key frames, the mean square error function criterion (MSE) of SAR image target block fusion is obtained. The calculation formula is as follows:

$$MSE(e_1, e_2) = \sum_{x=1}^{N_1} \sum_{y=2}^{N_2} \frac{f_i(x, y) n_i}{f_{i-1}(x + e_1, y + e_2) (l_x, l_y)} \quad (10)$$

Where,  $N_1 \times N_2$  is the high-frequency band coding bandwidth distribution of SAR image target extracted by video codec framework,  $(l_x, l_y)$  is the block fusion vector of SAR image template, and  $f_i(x, y)$  and  $f_{i-1}(x + e_1, y + e_2)$  represent the pixels of current frame and reference frame of SAR image target, respectively.

1	2	6	7
3	5	8	13
4	9	12	14
10	11	15	16

Fig. 2. Block Matching Structure Model of SAR Image Target.

The image pixel spatial fusion matching technology is used to realize the statistical analysis of weak and small target information in SAR image at each scale, and the statistical feature  $G(c_1, c_2)$  is:

$$G(c_1, c_2) = \frac{z \cdot \text{Length}(C) \text{MSE}(e_1, e_2)}{h \cdot \text{Area}(\text{inside}(C))} + \theta_1 |I - c_1| + \theta_2 |I - c_2| \quad (11)$$

Where,  $c_1$  and  $c_2$  represent the gray coefficient and brightness coefficient of SAR image target respectively, and  $z$ ,  $h$ ,  $\theta_1$  and  $\theta_2$  represent the sparsity feature distribution function, both of which are constants greater than 0. The key frame detection method is used to analyze the key frames in the target statistical feature. The calculation formula of frequency band division  $T$  of pixel key frames is as follows:

$$T = \frac{G(c_1, c_2)}{L_{\text{low}} \times L} \sum_{l \in \text{Lowfreq. } l=1}^L (E_k'(l) - E_k(l))^2 \quad (12)$$

Where,  $E_k'$  is the low-frequency band part of the similarity information fusion feature component,  $E_k$  is the low-frequency band part of the SAR image target pixel space,  $L$  is the number of DCT blocks of the SAR image target in each frame, and  $L_{\text{low}}$  is the number of low-frequency bands.

#### E. Dim and Small Target Recognition in SAR Image based on Intra Coding Function

According to the above obtained pixel key frame frequency band division results, track and analyze the weak and small targets in SAR image, mainly by constructing the intra coding function to realize tracking, and identify the weak and small targets in SAR image according to the tracking results. Firstly, the trajectory tracking function of each frame in the weak and small target area of SAR image is constructed, which is defined as follows:

$$v(g) = Tu^{-1}(u(1) - u(c(x))) \quad (13)$$

Where,  $c(g)$  is the neighborhood gray function of SAR image target, and  $u(\cdot)$  represents the trajectory tracking target function in key frame coding mode, which meets  $u: [0, 1] \rightarrow [0, 1]$ . Thus, the information feature quantity of weak and small targets in SAR image is extracted, and the expression of association rule coefficient  $P(w)$  of targets in SAR image is obtained as follows:

$$F(w) = \frac{\exp\left\{-\delta \sum_{I \in C} T_I(w)\right\}}{\sum_w \exp\left\{-\delta \sum_{I \in C} T_I(w)\right\}} v(g) \quad (14)$$

Where,  $\sum_{I \in C} T_I(w)$  is the total number of boundary pixels of weak and small targets in SAR image, and  $I$  is the spatial region neighborhood group of weak and small targets in SAR image. Then, the weak and small target recognition model of SAR image is constructed based on the intra coding function:

$$J = \sum_{k=1}^n \phi_r d(X_k, v_i) + F(w) \sum_{k=1}^n \phi_r d(\varepsilon, v_i) \quad (15)$$

Where,  $\phi_r$  represents a neighborhood of the target image collected in the  $r$ -th subband;  $\varepsilon$  represents the low frequency band part of the average similarity information fusion feature component. According to the sparse prior representation results, the high-resolution prediction value at frame  $m$  ( $x, y$ ) of the weak target  $F_m(x, y)$  in the SAR image is obtained, and the weak target recognition results of the super-pixel SAR image are obtained to improve the detection and recognition ability of the weak target in the SAR image.

### III. SIMULATION EXPERIMENT ANALYSIS

In order to verify the effectiveness of the weak and small target recognition method in SAR image based on machine vision in practical application, a simulation experiment is carried out by Vega software. Radar works is an important module in Vega software. It can produce real-time imaging radar simulation images based on physical mechanism. The operating environment of imaging radar is a comprehensive simulation environment composed of natural background, cultural characteristics and dynamic targets. One of the characteristics of radar works is that radar works runs in the same synthetic environment as Vega and sensor vision, and jointly provides fully correlated output windows and radar images. Before using the radar works module. It is necessary to set Vega basic modules (system configuration, window, channel, object, observer, motion mode, environmental effect, etc.), and then set the parameters of radar works module (radar type, resolution, frequency band, polarization mode, RCS range, surface side length, motion compensation mode, speckle noise level, carrier speed, image output mode, etc.) are set. The interface of radar works module in the visualization window Lyn X of Vega software is shown in Fig. 3.

The experimental parameter settings are shown in Table I.

Radar types in radar works module include real-time beam ground mapping, multi killer beam sharpening and synthetic aperture radar. The research in this chapter is only aimed at SAR. Radarworks supports six radar operating bands: K, Ku, x, C, s and l, and four polarization modes: VV, VH, HV and HH. The specific imaging parameter plan of SAR is shown in Fig. 4.

In this paper, a total of 8 SAR images are obtained through MSTAR database, including 26 ship targets, which are weak and small targets. They are used as the experimental samples for simulation test. The SAR image is shown in Fig. 5.

First, the algorithm proposed in the study is tested and applied for verification, that is, the loss value results of the algorithm with or without dimension reduction processing are counted. The results are shown in Fig. 6. The results of Fig. 6 show that before the algorithm is improved, the curve changes between the test loss value and the training loss value have a large difference, and when the epoch exceeds 200, the error trend of the two curves is large. However, the loss trend of the algorithm after dimension reduction is basically unchanged, and the difference affected by the value of epoch is small, indicating that the data loss has been improved.

Then, in order to further verify the effectiveness of this method, the SAR image dim target recognition method proposed in this paper, the multi azimuth SAR image target recognition method based on depth learning proposed in reference [12] and the SAR image target recognition method

based on 2dpca scn regularization proposed in reference [13] are used to identify dim targets in SAR images, and the recognition accuracy of the three methods is verified, The result data is compared from two aspects of recognition rate and false detection rate, and the comparison results are shown in Fig. 7.

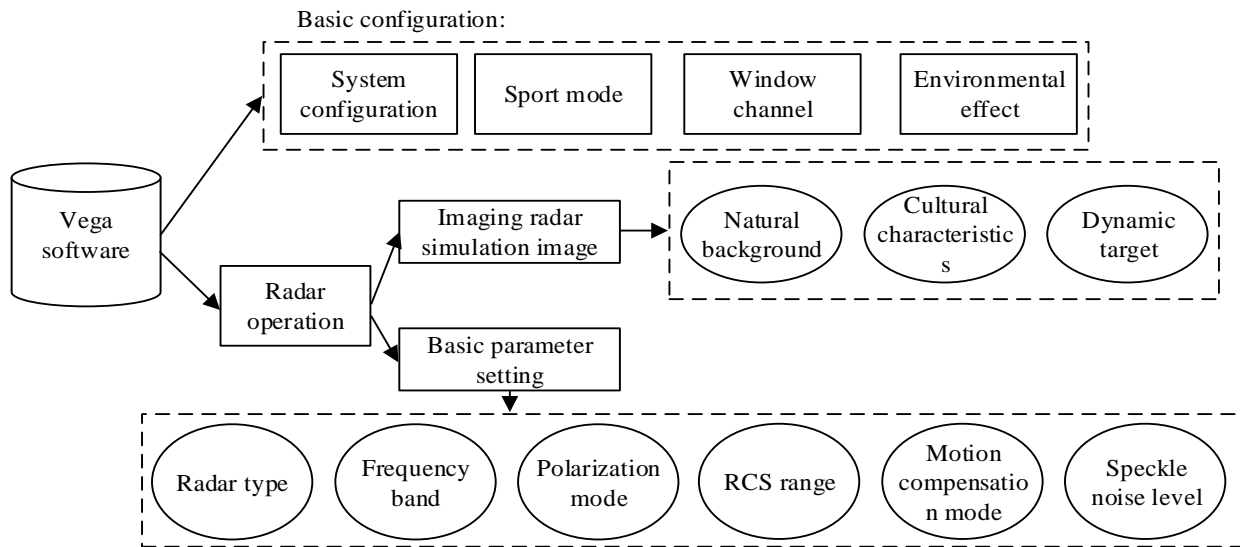


Fig. 3. Visual Operation Diagram of Vega Software.

TABLE I. EXPERIMENTAL PARAMETER SETTING

	Parameter
Frame frequency of visual sampling	12khz
Pixel set of image feature distribution	120
Identification interval	1.5ms
Regional pixel distribution	200*200
Action characteristic decomposition coefficient	0.46, 0.43

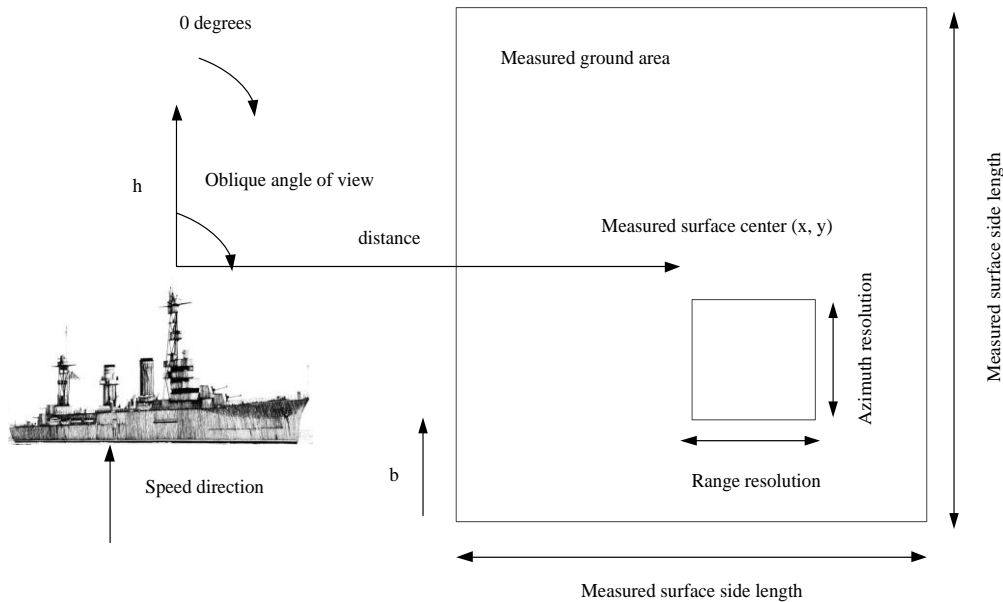


Fig. 4. Plan View of SAR Specific Imaging Parameters.

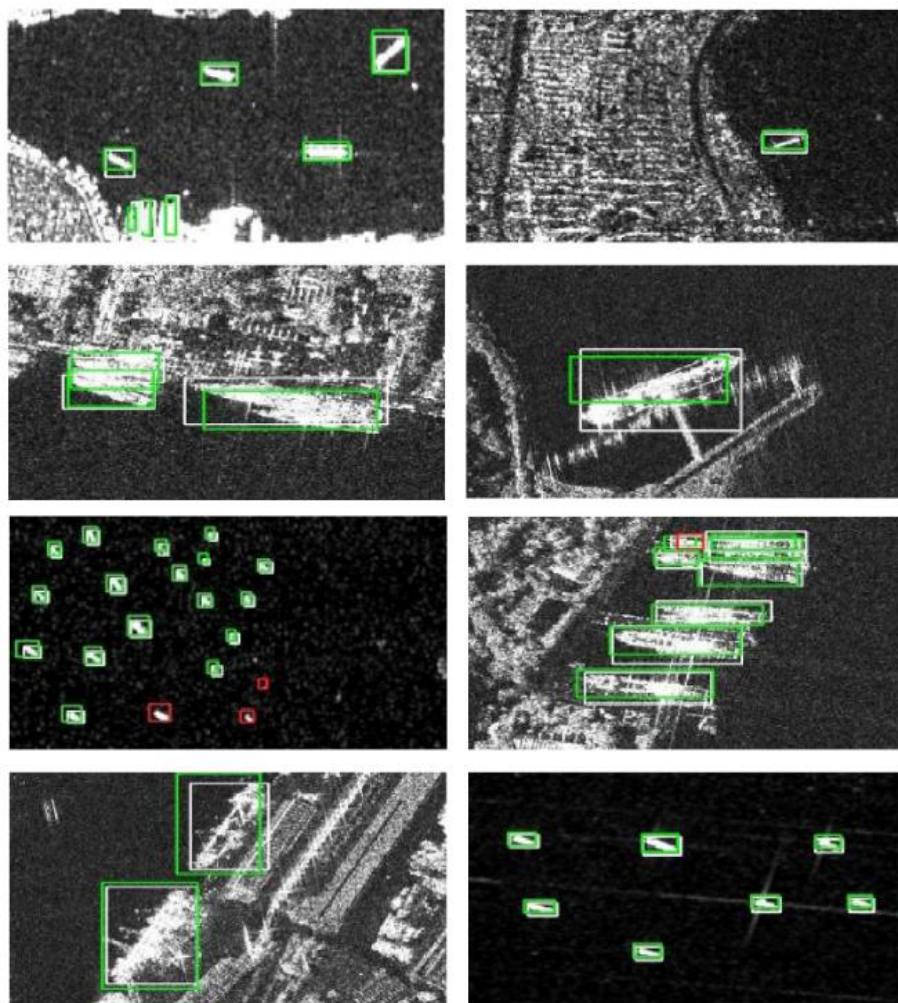


Fig. 5. SAR Image Samples.

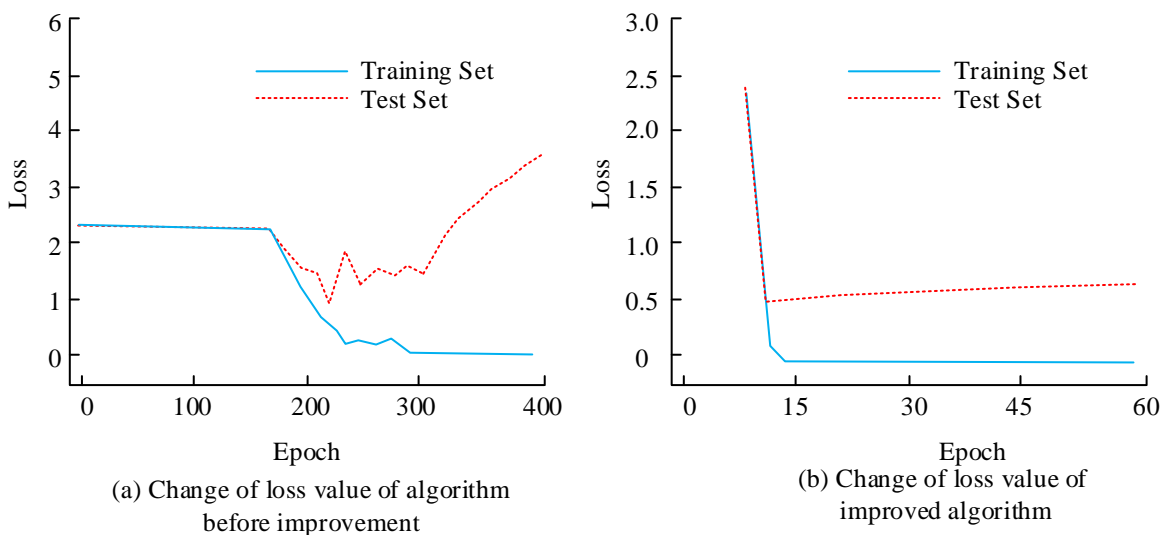


Fig. 6. Comparison Results of Weak and Small Target Recognition Accuracy in SAR Images of Three Methods.

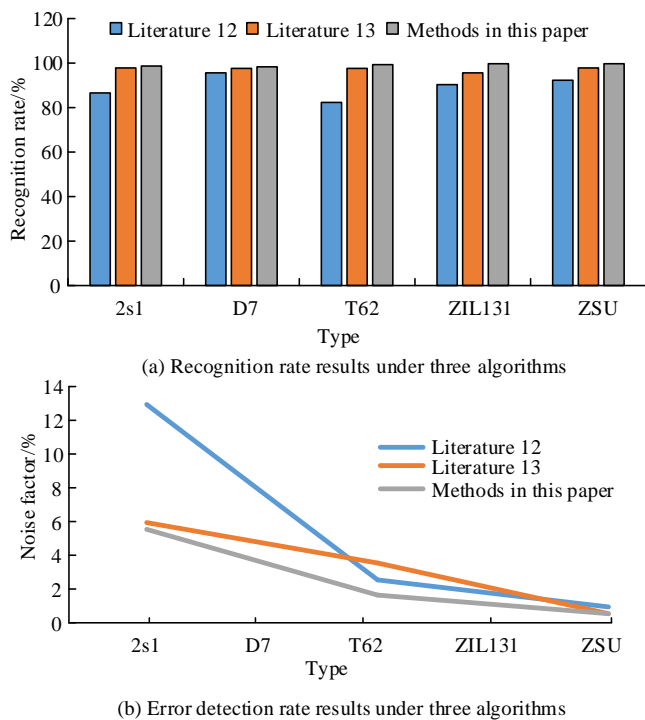


Fig. 7. Comparison Results of Recognition Rate and False Detection Rate.

Fig. 7 compares the algorithms from the two aspects of target recognition rate and false detection rate. It can be seen from Fig. 7(a) that the algorithms proposed in the study show good recognition rates in different types of test sets, with the recognition rates basically above 98%, and the average accuracy reaching 99%, far higher than the recognition rates shown in [12] and [13]. It can be seen from Fig. 7(b) that the false detection rate of the proposed algorithm is lower than that of the other two algorithms, with an average of less than 0.9%, while the average false detection rate of literature [12] and literature [13] is 6.9% and 2.1%. The above results show that the SAR image recognition method based on machine vision has good recognition rate and false detection rate, and the overall performance is good.

The above experimental results show that the algorithm proposed in this paper has a good recognition rate in weak target recognition, and effectively realizes the processing of image data information. Compare the research results with literature [27]. Literature [27] proposed to improve the accuracy of target recognition in SAR images by improving CFAR algorithm and operation algorithm. Its expansion and research of target data proposed that the dimension reduction of data considered the limited factors in target recognition process. The algorithm proposed in literature [27] has good inspection accuracy, which is similar to the algorithm results proposed in the research, which shows that the machine vision algorithm can improve the application performance of the algorithm.

#### IV. CONCLUSION

SAR is an active microwave imaging sensor. Its imaging principle is to install SAR on the radar platform, transmit electromagnetic waves regularly with the movement of the radar platform, and convert the received ground backscatter

signal into SAR image information. SAR uses pulse compression technology to obtain high resolution, can work all day and all weather, has multi band and multi polarization working mode, has certain penetration ability to soil, vegetation, clouds, etc., and can continuously image the observation area, so as to identify obstacles hidden in trees and forests. SAR is widely used in military, geographical and national economic issues. Automatic target recognition (ATR) of SAR image is a key research direction of SAR image interpretation. SAR image target recognition is the combination of SAR image manual interpretation and computer automatic recognition processing. Its working process can be described as: finding out the region of interest in SAR image, and then classifying each region of interest to determine its category. Therefore, this paper proposes a weak and small target recognition method based on machine vision in SAR image, and the effectiveness of this method is verified by simulation experiments. The algorithm proposed in the study considers the impact of linear characteristics on data when extracting image data information, so it uses principal component analysis to reduce the dimensions of data. Different from the improvement of previous improved algorithms in large dimensions, the study pays more attention to multi-dimensional consideration of data information.

In view of the limited ability and short time, in addition to making some progress, there are still many technical problems to be solved and improved, and the practical feasibility of the design scheme needs to be tested and corrected in the project. In order to obtain ideal detection results, there are still many aspects of technology to be studied.

1) This paper mainly studies the weak and small target recognition in SAR image under static background. In fact, the background may be dynamic. Further research is needed to realize the weak and small target recognition in SAR image under dynamic background.

2) In this paper, the target is not tracked after the recognition of weak and small targets in SAR images. Tracking the target is essential to realize the real-time monitoring or attack of the target. Further research is needed in this regard.

3) The engineering implementation of the weak and small target recognition method in SAR image studied in this paper needs further research, especially the setting of recognition method parameters. At present, the setting of parameters is mostly based on experience. After further research, artificial intelligence method can be used to set parameters.

#### REFERENCES

- [1] Z. Q. Su and Q.M. Ma, "Fuzzy image recognition of traffic signs based on convolution neural network," computer simulation, vol. 37, pp. 117-120,198, 2020.
- [2] B. Ding and G. Wen, "A Region Matching Approach based on 3-D Scattering Center Model with Application to SAR Target Recognition," IEEE Sensors Journal, vol. 11, pp. 1-7, 2018.
- [3] S. Zhang, Q. Cheng and D. Chen, et al., "Image Target Recognition Model of Multichannel Structure Convolutional Neural Network Training Automatic Encoder," IEEE Access, vol. 99, pp. 11-15, 2020.
- [4] J Chen, "Target Recognition of Basketball Sports Image Based on Embedded System and Internet of Things," Microprocessors and Microsystems, vol. 82, pp. 103918-103918, 2021.
- [5] X Wang, K. Zhang and J. Yan, et al., "Infrared Image Complexity Metric



- for Automatic Target Recognition Based on Neural Network and Traditional Approach Fusion,” *Arabian Journal for Science and Engineering, Section A, Sciences*, vol. 45, pp. 3245-3255, 2020.
- [6] Zhai J A, Dong G B, Chen F B, et al. “A Deep Learning Fusion Recognition Method Based on SAR Image Data,” *Procedia Computer Science*, vol. 147, pp. 533-541, 2019.
- [7] G. Chen, L. Wang, M. M. Kamruzzaman “Spectral classification of ecological spatial polarization SAR image based on target decomposition algorithm and machine learning,” *Neural Computing and Applications*, vol. 32, pp. 5449-5460, 2020.
- [8] He Z, Xiao H, Tian Z. “Multi-View Tensor Sparse Representation Model for SAR Target Recognition,” *IEEE Access*, vol.7, pp. 48256-48265, 2019.
- [9] R. Cao, Y. Wang, B. Zhao, et al. “Ship Target Imaging in Airborne SAR System Based on Automatic Image Segmentation and ISAR Technique”. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol.99, pp.1-1, 2021.
- [10] C. Wang, P. Huang, Y. Hu, “SAR target recognition method based on adaptive kernel dictionary learning,” *Dianbo Kexue Xuebao/Chinese Journal of Radio Science*, vol.34, pp. 60-64, 2019.
- [11] Q. Xu, W. Li, R. Zhan, et al. “Improved algorithm for SAR target recognition based on the convolutional neural network,” *Journal of Xidian University*, vol. 45, pp.177-183 2018.
- [12] H. Zou, Y. Lin and W. Hong, “Research on multi azimuth SAR image target recognition using depth learning,” *Signal processing*, vol. 34, pp. 513-522, 2018.
- [13] Y. P. Wang, Y. B. Zhang, Y. Li, et al., “SAR image target recognition method based on 2dpca-sc regularization,” *Signal processing*, vol. 35, pp. 802-808, 2019.
- [14] W. Shen and P. Shi, “Target recognition method of SAR image based on random weighted fusion of single signal,” *Journal of electronic measurement and instrumentation*, vol. 32, pp. 181-187, 2020.
- [15] P. P. Zhang, H.B. Luo, M.R. Ju, et al., “An improved capsule and its application in SAR image target recognition,” *Infrared and laser engineering*, vol. 49, pp. 195-202, 2020.
- [16] Y. R. Cho, S. Shin, S.H. Yim, et al., “Multistage Fusion with Dissimilarity Regularization for SAR/IR Target Recognition,” *Quality Control, Transactions*, vol. 7, pp. 728-740, 2019.
- [17] Y. Ma and Y. Li, “Millimeter-Wave InSAR Target Recognition with Deep Convolutional Neural Network,” *IEICE Transactions on Information and Systems*, vol. 102, pp. 655-658, 2019.
- [18] Z. He, H. Xiao, C. Gao, et al., “Fusion of Sparse Model Based on Randomly Erased Image for SAR Occluded Target Recognition,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 99, pp. 1-16, 2020.
- [19] Q. Chang and Z. Xiong, “Vision-aware target recognition toward autonomous robot by Kinect sensors,” *Signal Processing Image Communication*, vol. 84, pp. 115810-115817, 2020.
- [20] H. Ye, L. Zhang and D. Zhang, “Non-imaging target recognition algorithm based on projection matrix and image Euclidean distance by computational ghost imaging,” *Optics & Laser Technology*, vol. 137, pp. 106-117, 2021.
- [21] M. Chang, X. You and Z. Cao, “Bidimensional Empirical Mode Decomposition for SAR Image Feature Extraction with Application to Target Recognition,” *IEEE Access*, vol. 99, pp. 21-32, 2019.
- [22] B. Xue, W. Yi, F. Jing, et al., “Complex ISAR target recognition using deep adaptive learning,” *Engineering Applications of Artificial Intelligence*, vol. 97, pp. 104025-104030, 2021.
- [23] J. Wang, J. Liu, P. Ren, et al., “A SAR Target Recognition Based on Guided Reconstruction and Weighted Norm-Constrained Deep Belief Network,” *IEEE Access*, vol. 8, pp. 181712-181722, 2020.
- [24] X. Yang, X. Nan and B. Song, “D2N4: A Discriminative Deep Nearest Neighbor Neural Network for Few-Shot Space Target Recognition,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 99, pp. 1-10, 2020.
- [25] P. Bolourchi, M. Moradi, H. Demirel, et al., “Improved SAR target recognition by selecting moment methods based on Fisher score,” *Signal, Image and Video Processing*, vol. 14, pp. 39-47, 2020.
- [26] X. Bai, X. Zhou, F. Zhang, et al., “Robust Pol-ISAR Target Recognition Based on ST-MC-DCNN,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 99, pp. 1-16, 2019.
- [27] Y. Wang, T. Guo, “Improved Ship Target Detection Accuracy in SAR Image Based on Modified CFAR Algorithm,” *Journal of Harbin Institute of Technology (New Series)*, vol.25, pp. 18-23, 2018.

# Information Classification Algorithm based on Project-based Learning Data-driven and Stochastic Grid

Xiaomei Qin, Wenlan Zhang\*

School of Education, Shaanxi Normal University, Xi'an Shaanxi 710062, China

**Abstract**—The adaptive partitioning algorithm of information set in simulation laboratory based on project-based learning data-driven and random grid is studied to effectively preprocess the information set and improve the adaptive partitioning effect of the information set. Using the improved fuzzy C-means clustering algorithm driven by project-based learning data, the fuzzy partition of information set in simulation laboratory is carried out to complete preprocessing of information set; The pre-processing information set space is roughly divided by the grid partitioning algorithm based on the data histogram; A random mesh generation algorithm based on uniformity is used to finely divide the coarse mesh cells; Taking the representative points of grid cells as the clustering center, the pre-processing information set is clustered by the density peak clustering algorithm to complete the adaptive partitioning of the information set in simulation laboratory. Experimental results show that this algorithm can effectively preprocess and adaptively partition the information set of simulation laboratory; For different dimension information sets, the evaluation index values of Rand index, Purity, standard mutual information, interval and Dunn index of the algorithm are all high, and the evaluation index values of compactness and Davidson's banding index are all low, so the algorithm has a high accuracy of adaptive partitioning of information sets.

**Keywords**—Adaptive partitioning; data driven; information set; project-based learning; random grid; simulation laboratory

## I. INTRODUCTION

Simulation experiments exist in the simulation laboratory, and the experimental environment, objects and equipment are provided by the simulation laboratory [1]. The simulation laboratory is a kind of simulation experiment environment supported by computer software and hardware technology and realized by software development tools [2]. By developing a series of simulation experiment components to simulate and reproduce the experimental environment, experimental equipment and experimental process, the experimenter can get rid of the bondage of the actual experimental conditions, feel the experimental information interactively, and realize the experimental process in a near real way under more convenient and fast conditions [3]. In the simulation laboratory, the experimental objects and equipment are either reproduced or simulated vividly, and the experimental process is completely controlled by the experimenter [4]. The simulation laboratory uses the powerful computing processing ability of the computer, with the help of graphics / images, simulation and virtual reality technologies, and has rich interface information,

friendly interaction ability and powerful data processing function. In addition, it is well compatible with various external devices, multi-media and Internet, forming a wonderful simulation experiment world. Project-based learning is a new teaching mode, which mainly focuses on the core concepts and principles of the discipline, and emphasizes that learners can solve practical problems, participate in some exploratory activities and other meaningful learning tasks [5]. In this process, learners learn independently and construct the meaning of the learned knowledge through practical operation, and apply it in the simulation laboratory. It can effectively improve learners' interest in simulation experiments and learning quality. Once the project-based learning simulation laboratory is established, it is a shared resource. Although it is not limited by the site and time, it greatly saves material resources, but it will generate a large number of project-based learning data and increase the difficulty of data search. The best way to solve this problem is to study a partitioning algorithm, which divides all information into several categories through the partitioning algorithm to facilitate information search. For example, Sutagundar, A., et al. proposed to use the function of sensor cloud and fog calculation to divide in a better way and minimize the delay problem. Using random forest classifier and genetic algorithm to divide the information and introducing Agent paradigm, not only saved the energy of physical sensor nodes, but also could quickly analyze and divide the information on the fog server. The results show that the algorithm has better effect in partition accuracy, delay and energy consumption [6]; Flisar, J., et al. used DBpedia ontology knowledge base to divide information. This algorithm can effectively divide information sets [7], but it cannot adapt to all data distribution well. The partitioning algorithm boundary is easily discarded as noise points, resulting in low partitioning accuracy. In order to achieve efficient, high-precision and high-speed information partition, Zheng, T., et al. designed a partitioning algorithm with low power consumption and high performance. In the training phase, a laser radar was simulated to collect the laser radar information, and then the designed neural network was trained using the information and the corresponding tags. In the testing phase, the new information was first divided into simple units using the range transformation watershed method, then, the trained neural network was used to divide the information. The average recall rate of the information partitioning of the algorithm was 0.965 and the average precision rate was 0.943 [8], but its complexity was high and it was difficult to meet the real-time requirements, and there were limitations in the processing of

\*Corresponding Author.

multidimensional information sets. The algorithm is susceptible to noise and the effect of boundary processing is not ideal. Data driving can be broadly defined as using the online or offline data of the system to realize various expected functions of the system such as data-based preprocessing, evaluation, scheduling, monitoring, diagnosis, decision-making and partitioning algorithm [9]. The random grid density clustering algorithm is insensitive to the input data and can adapt to different data distributions. Moreover, the algorithm has low complexity and fast calculation speed. It is suitable for partitioning different dimensional information and is not sensitive to the impact of noise [10].

In order to improve the precision of adaptive partitioning of information set, an adaptive partitioning algorithm of information set in simulation laboratory based on project-based learning data-driven and random grid is studied. First of all, the paper uses the fuzzy C-means clustering algorithm to carry out the fuzzy division of information, and preliminarily classifies the information. On this basis, the paper further uses the data histogram grid division algorithm to further roughly divide the information space, strengthen the classification of information, and avoid the problem of insufficient classification caused by the strong correlation between information; Then, the information is further clustered with the density peak clustering algorithm, which strengthens the aggregation of similar information and completes the information division. Finally, the experimental results show that this research method has the ability of adaptive partition, high precision of partition, and good overall application.

## II. ADAPTIVE PARTITIONING ALGORITHM OF INFORMATION SET IN SIMULATION LABORATORY

### A. Information Set Preprocessing of Simulation Laboratory by using the Improved Fuzzy C-means Clustering Algorithm based on Project-based Learning Data-driven

In the simulation laboratory, the project-based learning method is applied to carry out simulation experiments, so that students can explore the selected simulation experiment projects according to their own interests and learning needs. This learning activity emphasizes the students' hands-on operation and comprehensive application of various knowledge achievements, which can improve the learning quality of each simulation experiment project stage and make students more interested in the learning process of simulation experiments. In the process of simulation experiments using project-based learning, a large number of project-based learning data will be generated in the information of the simulation laboratory. In order to further improve the learning quality of each project stage in the simulation experiment process, it is necessary to self-adapt the project-based learning data in the information set of the simulation laboratory [11], accelerate the efficiency of students viewing the project-based learning data in the information set of the simulation laboratory, and facilitate the management of the information set of the simulation laboratory. In order to improve the adaptive partitioning effect of the project-based learning data in the information set of the simulation laboratory [12], it is necessary to obtain the fuzzy version of the project-based learning data in the original information set of the simulation laboratory (composed of

fuzzy attributes and fuzzy partitions). Fuzzy C-means clustering algorithm (FCM) generates fuzzy partitions by defining the value of fuzzy membership function ( $\mu$ ), which can solve the problem of hard boundary value caused by sharp partitions and protect the item learning data in the original information set. In addition, the fuzzy partition has obvious semantic relevance, which can well solve the inherent uncertainty of numerical data in the project-based learning data in the information set of the simulation laboratory.

An expression called fuzzy entropy is used as the cost function of the objective function of FCM algorithm. The definition of fuzzy entropy is roughly the same as that of information entropy, which is more suitable for fuzzy clustering analysis [13]. The function of fuzzy entropy can be defined as:

$$E(x) = - \sum_{i=1}^c \sum_{j=1}^n m_{ij} \mu_{ij} \ln m_{ij} \mu_{ij} \quad (1)$$

Wherein, the project-based learning data in the information set of simulation laboratory is  $x$ ; The fuzzy membership degree of the  $i$ -th attribute of the project-based learning data in the  $i$ -th information set of simulation laboratory is  $\mu_{ij}$ ;  $m$  is a weighted index, whose physical meaning is the fuzziness degree constant; The number of project-based learning data in the simulation laboratory information set is  $c$ ; The number of fuzzy attributes of the project-based learning data in the information set of the simulation laboratory is  $n$ .

The objective function of FCM can be defined as:

$$\begin{aligned} \min J(\mu, v) &= \sum_{i=1}^c \sum_{j=1}^n m_{ij} d_{ij}^2 \\ \text{s.t. } \sum_{i=1}^c \mu_{ij} &= 1 \end{aligned} \quad (2)$$

Where, the objective function of the sum of error squares is  $\min J(\mu, v)$ , and its value reflects the degree of compactness within the class under a certain difference definition [14]. The smaller the value of  $\min J(\mu, v)$  is, the tighter the clustering is; The clustering center of project-based learning data in the information set of simulation laboratory is  $v$ ; The Euclidean distance is  $d_{ij}^2$ .

Taking information entropy as the cost function, the minimum error square sum objective function of FCM algorithm is used to solve the problem of fuzzy partition of project-based learning data in the information set of simulation laboratory. The objective function of the improved FCM algorithm is:

$$L(\mu, v, \alpha) = \min J(\mu, v) + \alpha E(x) \quad (3)$$

Where the Lagrange multiplier is  $\alpha$ .

Substituting formula (1) and formula (2) in formula (3) can obtain:

$$L(\mu, v, \alpha) = \sum_{i=1}^c \sum_{j=1}^n m \mu_{ij} d_{ij}^2 - \sum_{i=1}^c \sum_{j=1}^n \alpha_j m \mu_{ij} \ln m \mu_{ij}^m \quad (4)$$

The Lagrange multiplier of the j-th attribute of the project-based learning data in the information set of the simulation laboratory is  $\alpha_j$ .

If  $\frac{\partial L}{\partial \mu_{ij}} = 0$  in formula (4), the following can be obtained:

$$\mu_{ij} = \exp\left(\frac{d_{ij}^2}{m \alpha_j} - \frac{1}{m}\right) \quad (5)$$

Since  $\sum_{i=1}^c \mu_{ij} = 1$ , the finishing formula (5) can be obtained:

$$\alpha_j = \frac{1}{\sum_{k=1}^c d_{kj}^2} \quad (6)$$

Wherein the attribute number of the project-based learning data in the information set of simulation laboratory is k.

Substituting formula (5) and formula (6) into formula (4), it can get:

$$L(\mu, v, \alpha) = \frac{m \mu_{ij}}{\sum_{k=1}^c d_{kj}^2} \quad (7)$$

It can be seen from formula (7) that in order to obtain the optimal value of  $\min J(\mu, v)$ , it is also necessary to process the project-based learning data in the information set of simulation laboratory [15]. By analyzing the physical meaning

of  $\frac{1}{\sum_{k=1}^c d_{kj}^2}$ , it can be seen that in fact it should be the distribution characteristics of project-based learning data in the information set of the simulation laboratory, representing a distribution characteristic of project-based learning data in the data space [16]. That is to say, taking information entropy as the cost function of the objective function, the objective function of the FCM algorithm needs to obtain the minimum value. In addition to the membership degree [17], the actual distribution characteristics of the project-based learning data in the data space in each cluster must also be considered. For the

convenience of later expression and calculation,  $\delta_j = \frac{1}{\sum_{k=1}^c d_{kj}^2}$  is specially used, that is,  $\delta_j$  is used to represent this distribution characteristic.

According to the above analysis, the objective function is adjusted accordingly to obtain:

$$J(\mu, v, \delta) = \sum_{i=1}^c \sum_{j=1}^n m \delta_j m \mu_{ij} d_{ij}^2$$

$$s.t. \sum_{i=1}^c \mu_{ij} = 1 \quad (8)$$

Then, according to the Lagrange function method, it can get:

$$J(\mu, v, \delta, \alpha) = \min \sum_{i=1}^c \sum_{j=1}^n m \delta_j m \mu_{ij} d_{ij}^2 + \sum_j \alpha_j \left(1 - \sum_{i=1}^c \mu_{ij}\right) \quad (9)$$

Let  $\frac{\partial L}{\partial v_i} = 0$ , then:

$$v_i = \frac{\sum_{j=1}^n m \delta_j m \mu_{ij} x_j}{\sum_{j=1}^n m \delta_j m \mu_{ij}} \quad (10)$$

Wherein the project-based learning data in the information set of simulation laboratory of the j-th attribute is  $x_j$ ; The clustering center of the project-based learning data in the i-th simulation laboratory information set is  $v_i$ .

Let  $\frac{\partial L}{\partial \mu_{ij}} = 0$ , then:

$$\mu_{ij} = \left(\frac{\alpha_j}{m \delta_j d_{ij}^2}\right)^{\frac{1}{m-1}} \quad (11)$$

The fuzzy partition of the project-based learning data in the information set of the simulation laboratory is generated by iteratively optimizing the objective function formula (8) by

updating the values of the membership function  $\mu_{ij}$  and the cluster center  $v_i$ .

In order to improve the fuzzy partition effect of the project-based learning data in the information set of the simulation laboratory, the fuzzy membership function of the FCM algorithm is determined by using the project-based learning data-driven method. The generation process of fuzzy partition of project-based learning data in the information set of simulation laboratory based on project-based learning data driven FCM is as follows:

It is defined that the project-based learning data in the information set of the simulation laboratory is  $D = \{x_1, x_2, \dots, x_c\}$ ,  $x_1, x_2, \dots, x_c$  is the data of different

numerical attributes, and D is the set of Boolean attributes and numerical attributes; The numeric attribute set is defined as  $A = \{q_1, q_2, \dots, q_r\}$ ; The fuzzy partition is defined as  $P = \{P_1, P_2, \dots, P_r\}$ , the fuzzy partition is the result of clustering the numerical attribute set A by FCM algorithm, and  $P_r = \{f_1, f_2, \dots, f_w\}$  is the fuzzy clustering partition set of the numerical attribute  $q_r$ .

The improved FCM algorithm is used to cluster the data in the numerical attribute set of project-based learning data in each information set of simulation laboratory, and then the corresponding fuzzy partition is obtained. Each numerical attribute data has its own unique fuzzy membership function. This process is repeated several times until the fuzzy partition of all numeric attribute values is obtained.

The whole preprocessing process of the project-based learning data in the information set of the simulation laboratory includes two steps: the first step is to generate fuzzy partitions for the values of the numerical attributes of the project-based learning data in each information set of the simulation laboratory through the improved FCM algorithm; The second step is to further process the project-based learning data in the original information set of simulation laboratory, and then obtain its fuzzy Version (composed of fuzzy attributes and fuzzy partitions). Pre definition: the Boolean attribute value

data set is  $B = \{b_1, b_2, \dots, b_{m'}\}$ ; The definition attribute set is

$\hat{A} = B \cup A$ . The fuzzy version generation process of the project-based learning data in the original information set of simulation laboratory is as follows: scan the project-based learning data in the original information set of simulation laboratory, classify the data attribute values, put the project-based learning data belonging to the Boolean attribute in  $B = \{b_1, b_2, \dots, b_{m'}\}$ , and put the project-based learning data

belonging to the numerical attribute in  $A = \{q_1, q_2, \dots, q_r\}$ , and then perform classification calculation. When an item learning data is a Boolean attribute, its fuzzy membership function  $\mu = 1$  or  $\mu = 0$  can easily divide the fuzzy partition, and then obtain the fuzzy version of the Boolean attribute value [18]; When a project-based learning data is a numerical attribute, each numerical attribute in the project-based learning data D can be converted into a fuzzy record according to the fuzzy partition  $P_r$ . Each fuzzy record contains the fuzzy attribute of the project-based learning data and the corresponding fuzzy membership function.

Thus, after selecting the data attributes in the first step, an intermediate version of the project-based learning data  $D_1$  in the information set of simulation laboratory can be generated; Then,  $D_1$  is updated iteratively [19], until all the attribute data in the attribute set  $\hat{A}$  are processed, and then the fuzzy version

of the original project-based learning data will be obtained. Therefore, by applying the FCM preprocessing technology driven by project-based learning data, any project-based learning data in the original information set of simulation laboratory with Boolean and numerical attributes can be transformed into a fuzzy set  $Q = \{\hat{x}_1, \hat{x}_2, \dots, \hat{x}_N\}$  with fuzzy records and fuzzy attributes.

### B. Adaptive Partitioning of Information Set in Simulation Laboratory based on Random Grid

1) *Random grid partitioning of information set in simulation laboratory*: The density peak clustering algorithm based on random grid partitioning is used to adaptively divide the fuzzy set  $Q = \{\hat{x}_1, \hat{x}_2, \dots, \hat{x}_N\}$  of the project-based learning data in the information set of simulation laboratory obtained in Section 2.1.

In order to better adapt to the best partitioning of the fuzzy sets of project-based learning data in different situations, first, the fuzzy sets of project-based learning data are roughly divided by the data histogram, and then the random grid is divided by the grid uniformity.

In order to adapt to the adaptive partitioning environment of fuzzy sets of project-based learning data, overcome the shortcomings of current grid division, and improve the efficiency and reliability of grid cell uniformity [20], firstly, rough partitioning of fuzzy sets of project-based learning data is carried out by using histograms. The steps are as follows:

Step 1: read in the fuzzy set  $Q = \{\hat{x}_1, \hat{x}_2, \dots, \hat{x}_N\}$  of the project-based learning data and normalize it. The normalization formula is:

$$\tilde{x} = \frac{\hat{x} - \hat{x}_{\min}}{\hat{x}_{\max} - \hat{x}_{\min}} \quad (12)$$

Wherein, the minimum and maximum values in the fuzzy set of project-based learning data in the information set of the simulation laboratory are  $\hat{x}_{\min}$  and  $\hat{x}_{\max}$ ; The item learning data in the fuzzy set after normalization processing is  $\tilde{x}$ .

Step 2: scan the fuzzy set Q of item learning data once, and draw the data histogram on each dimension within Q.

Step 3: according to the corresponding data histogram of each dimension, determine the high-density area and the low-density area of the fuzzy set of project-based learning data in the information set of the simulation laboratory, and divide the grid cells of the high-density area into fine ones and the grid cells of the low-density area into coarse ones. The specific operation is to use the number of data points contained in a grid to be equal to a given threshold  $\beta$  to divide, and the value of threshold  $\beta$  is related to the size of the entire fuzzy set Q.

Although the rough grid partitioning based on data histogram divides the regions with different densities of the project-based learning data fuzzy set in different scales, the

distribution of the project-based learning data of the grid cells with the same density may still be very different, so the boundary of the class cannot be effectively found and the class with arbitrary shape cannot be effectively found. Next, the random grid cells are finely divided by the uniformity index. The steps are as follows:

Step 1: calculate the uniformity of each random grid element obtained by rough division, and the formula is as follows:

$$S_l(u) = 1 - \left\{ \max \left( \frac{\phi_{i'}(u) - \eta_{i'}(u)}{\eta_{i'}(u)} \right) \right\} \quad (13)$$

Where  $\eta_{i'}(u)$  is used to represent the overall standard deviation of the  $i'$ -dimension of the random grid unit  $u$ ;  $\phi_{i'}(u)$  represents the standard deviation of the  $i'$ -dimensional project-based learning data samples of the random grid cell  $u$ , and  $S_l(u)$  represents the uniformity of the 1-dimensional grid cell  $u$ ;  $i' = 1, 2, \dots, l$ ; The closer the value of  $S_l(u)$  is to 1, the higher the uniformity of the random grid cells in the fuzzy set of item learning data is.

If  $S_l(u)$  satisfies the given threshold, i.e. it is a uniform mesh, it will not be divided; If the uniformity is less than the given threshold, i.e. non-uniform mesh, go to step 2.

Step 2: for each non-uniform grid, divide the random grid cell into two new grids of equal size along the worst dimension of  $S_l(u)$ , and judge whether the two new grids meet the stop criteria. If not, further divide them by the same method until the stop criteria are met.

Step 3: stop criteria. ① The grid element is a uniform grid; ② The grid cell is an empty grid; ③ The number of project-based learning data points in the grid is less than a given threshold.

2) Automatic selection of clustering center of information set in simulation laboratory: It is assumed that the center point of the grid unit  $u$  after random grid partitioning is  $g_u = (g_u^1, g_u^2, \dots, g_u^l)$  and  $u = 1, 2, \dots, K$ ;  $K$  is the total number of meshes; Where  $g_u$  represents the coordinate of the center point of the grid cell  $u$  in the 1-dimension, so the grid cell  $G_u$  can be expressed as:

$$G = \left[ \left( g_u^1 - \frac{side}{2}, g_u^1 + \frac{side}{2} \right), \left( g_u^2 - \frac{side}{2}, g_u^2 + \frac{side}{2} \right), \dots, \left( g_u^l - \frac{side}{2}, g_u^l + \frac{side}{2} \right) \right] \quad (14)$$

Wherein the side length of the grid cell is side.

The set of points in the grid cell  $u$  is  $Y = \{y_1, y_2, \dots, y_{h_u}\}$ , and  $h_u$  is the total number of project-based learning data points in the grid cell  $u$ , then the representative points of the grid cell are:

$$Y_u = \frac{\sum_{\hat{x}_i \in G_u} \hat{x}_i}{\lambda h_u} \quad (15)$$

Where the constant is  $\lambda$ ; The item learning data in the fuzzy set of the  $i$ -th information set in simulation laboratory after normalization is  $\hat{x}_i$ .

The local density of the grid cell's representative point  $Y_u$  is the number of project-based learning data points in the grid cell  $u$ , and the number of points in the grid cell  $u$  is:

$$h_u = \sum_{i=1}^N f'(\hat{x}_i, G_u) \quad (16)$$

Where, the function is  $f'(\cdot)$ ,  $f'(\hat{x}_i, G_u) = \begin{cases} 1 & g_u^l - \frac{side}{2} \leq \hat{x}_i < g_u^l + \frac{side}{2} \\ 0 & other \end{cases}$ , so the local density of

$Y_u$  is  $\rho_u = h_u$ .

The nearest distance between the grid cell's representative point  $Y_u$  and the higher density representative point  $Y_o$  is taken as the distance value of the grid cell's representative point  $Y_u$ , which is recorded as  $d'_u$ , and the formula is as follows:

$$d'_u = \min_{o: \rho_o > \rho_u} (D'_{ou}) \quad (17)$$

Wherein  $d'_u$  is the distance between the grid cell's representative point  $Y_u$  and the grid cell's representative point  $Y_o$ .

An improved adaptive method is designed to complete the automatic selection of cluster centers when the information set of simulation laboratory is adaptively divided, and the exact number of cluster centers is selected without manual intervention, so as to improve the accuracy of cluster center selection. The determination function is:

$$\rho_{C_i} - \sigma(\rho_i) \geq 0 \quad (18)$$

$$\frac{\xi_{C_i} - E'(\xi_i)}{2} \geq \varpi(\xi_i) \quad (19)$$



Wherein the density of the representative point  $C_i$  of the grid cell of the  $\hat{i}$ -th cluster center is  $\rho_{C_i}$ ; The mean value of the representative point density of all grid cells is  $\sigma(\rho_i)$ ; The minimum distance between the representative point of grid cell and the representative point of cluster center in the same cluster is  $\xi_{C_i}$ ; The expectation of all  $\xi_i$  is  $E'(\xi_i)$ . Formula (18) indicates that the local density value of the representative points of the grid cell is greater than the average value of the local density of all the representative points in the grid cell. This determination method satisfies the condition that the clustering centers of the project-based learning data in the information set of the simulation laboratory are often distributed in the relatively high density area in the density peak algorithm. The determination method of formula (19) satisfies the condition that the relative distance between the cluster centers is relatively long. Therefore, when the grid cell's representative point object meets the above two formula conditions, the grid cell's representative point is selected as the cluster center.

3) *Classification of project-based learning data points in the information set of simulation laboratory*: The nearest neighbor algorithm in the density peak clustering algorithm is used to classify the item learning data points in the fuzzy set of the remaining information set of simulation laboratory. After the selection of the representative points of the cluster center is completed, the remaining non cluster center representative points are classified into the class of the representative points closest to them and with local density greater than the point in  $\rho_i$ -descending order, and the data points in the project-based learning data in the fuzzy set of the original information set in simulation laboratory are assigned to the class of the representative points of the grid cells.

When the project-based learning data in the fuzzy set of the information set of the simulation laboratory is adaptively divided, the object of the boundary point is the representative point of the grid cell. Firstly, the set of boundary grid cell's representative points in the current cluster is calculated according to the density parameter  $\theta_c$ , to find the grid cell's representative points with the highest density in the boundary point set, and take the density of the representative points as the threshold to divide the core representative points and noise points, so as to reserve the representative points with the density greater than or equal to the density threshold as the core representative points in the cluster; The noise representative points in the current category that are smaller than the density threshold are removed, and the project-based learning data points in the grid where the noise representative points are located are also removed.

4) *Adaptive partitioning process of information set in simulation laboratory*: The specific steps of the adaptive partitioning of the information set in simulation laboratory are as follows:

Step 1: use the project-based learning data in Section 2.1 to drive FCM, preprocess the project-based learning data in the information set of the simulation laboratory, and obtain the fuzzy set of the project-based learning data;

Step 2: normalize the item learning data in the fuzzy set;

Step 3: roughly mesh the fuzzy set of project-based learning data according to the data histogram;

Step 4: perform random mesh refinement on the coarse divided mesh according to the uniformity to obtain several disjoint mesh elements;

Step 5: map the project-based learning data points to the corresponding grid cells, obtain the representative points of each grid cell from formulas (14) and (15), and count the number of project-based learning data points contained in each grid cell;

Step 6: calculate the local density  $\rho_u$  of the representative point  $Y_u$  of the grid cell according to formula (16);

Step 7: arrange the grid cell's representative points in reverse order according to  $\rho_i$ , and calculate the high density distance  $d'_u$  of each grid cell's representative point according to formula (17);

Step 8: adaptively determine the representative point of the cluster center by formula (18) and formula (19), classify it into the class of the grid representative point with the shortest distance and the local density greater than the point according to the p-descending order, and classify all data points in the project-based learning data in the original information set of simulation laboratory into the class of the grid representative point;

Step 9: calculate the boundary point set of the current class from the density parameter  $\theta_c$ , select the representative point with the highest density in the boundary point set, and use the density of the point as the threshold for dividing the core representative points and noise points of the current class, and eliminate the representative points in the current class that are smaller than the density threshold and other item learning data points in the grid cell where the representative points are located;

Step 10: return the final clustering result, that is, complete the adaptive partitioning of the simulation laboratory information set.

### III. EXPERIMENTAL ANALYSIS

Taking the sensor simulation laboratory of a university as the experimental object, the simulation laboratory realizes semi-automatic interactive control through animation, video or virtual reality technology. The structure of the simulation laboratory is shown in Fig. 1.

The simulation laboratory integrates the resources such as computers, instruments and equipment, tested points and their data into the network for sharing, and realizes the functions of

remote or remote testing, control, data acquisition, fault monitoring and on-site monitoring. The networked simulation instrument can obtain the measurement information from any place and at any time.

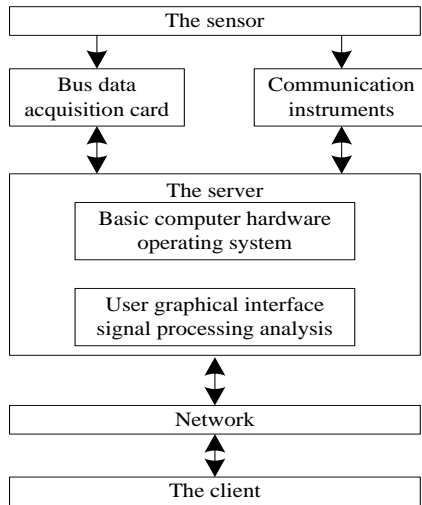


Fig. 1. Structure Diagram of Simulation Laboratory.

In the process of sensor simulation experiments in the project-based learning mode conducted by the university students in the simulation laboratory, tens of thousands of project-based learning data of different dimensions are generated, forming a high-dimensional information set in simulation laboratory and a low-dimensional information set in simulation experiment. Both information sets contain four types of project-based learning data, namely, carrier data, modulation data, excitation data and vibration data. Using the algorithm in this paper, the information sets of two simulation laboratories are divided adaptively, which proves that the algorithm in this paper has a good adaptive partitioning effect.

Taking the information set of low dimensional simulation experiment as an example, the information set is fuzzy partitioned by the algorithm in this paper, and the results of fuzzy partitioning are shown in Fig. 2.

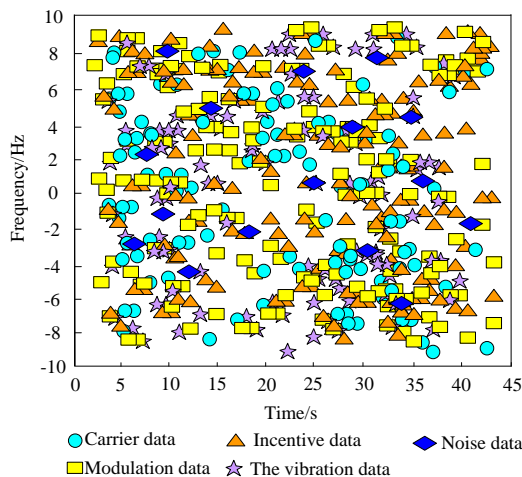


Fig. 2. Project-based Learning Data in the Original Low-dimensional Simulation Laboratory Information Set.

Using the quantitative attribute Income, the effect of fuzzy partition of the project-based learning data in the information of the low-dimensional simulation laboratory is analyzed. The value range of the Income attribute is  $Income \in [0, 9]$  (the unit is thousands). The number of fuzzy partitions of project-based learning data in the information set of simulation laboratory is defined as 4, which are "Around 1", "Around 3", "Around 5", "Around 7"; The analysis results of fuzzy partition effect of the algorithm in this paper are shown in Fig. 3.

According to Fig. 3, after preprocessing the project-based learning data in the original low-dimensional information set of simulation laboratory using the algorithm in this paper, the fuzzy partitions obtained are related to each other, that is, they have strong semantic relevance, and the curves of each fuzzy partition are relatively smooth, which solves the problem of "sharp partition", protects the boundary value, and thus protects the project-based learning data. The Income value corresponding to the highest value of the fuzzy membership degree of each fuzzy partition corresponds to the set Around value, which indicates that the algorithm in this paper has better fuzzy partition effect of the information set in simulation experiment, that is, better preprocessing effect of the information set.

The algorithm in this paper is used to adaptively partition the project-based learning data in the low-dimensional information set of simulation experiment. The adaptive partitioning results are shown in Fig. 4.

According to Fig. 4, the algorithm in this paper can effectively divide the project-based learning data in the information set of the simulation laboratory. After the rough division, the description between the project-based learning data is relatively fuzzy, and the noise points are not distinguished. After the random grid refinement, the boundary area between the project-based learning data is further refined, and the noise point data is effectively divided into isolated grid cells; According to the grid cells finely divided by random grid, four kinds of project-based learning data are obtained through clustering processing, and the noise data scattered outside the project-based learning data is better removed. Experimental results show that the proposed algorithm can effectively partition the information set of simulation laboratory.

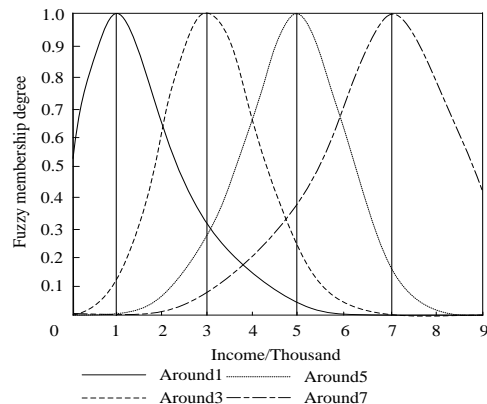


Fig. 3. Fuzzy Partition Effect.

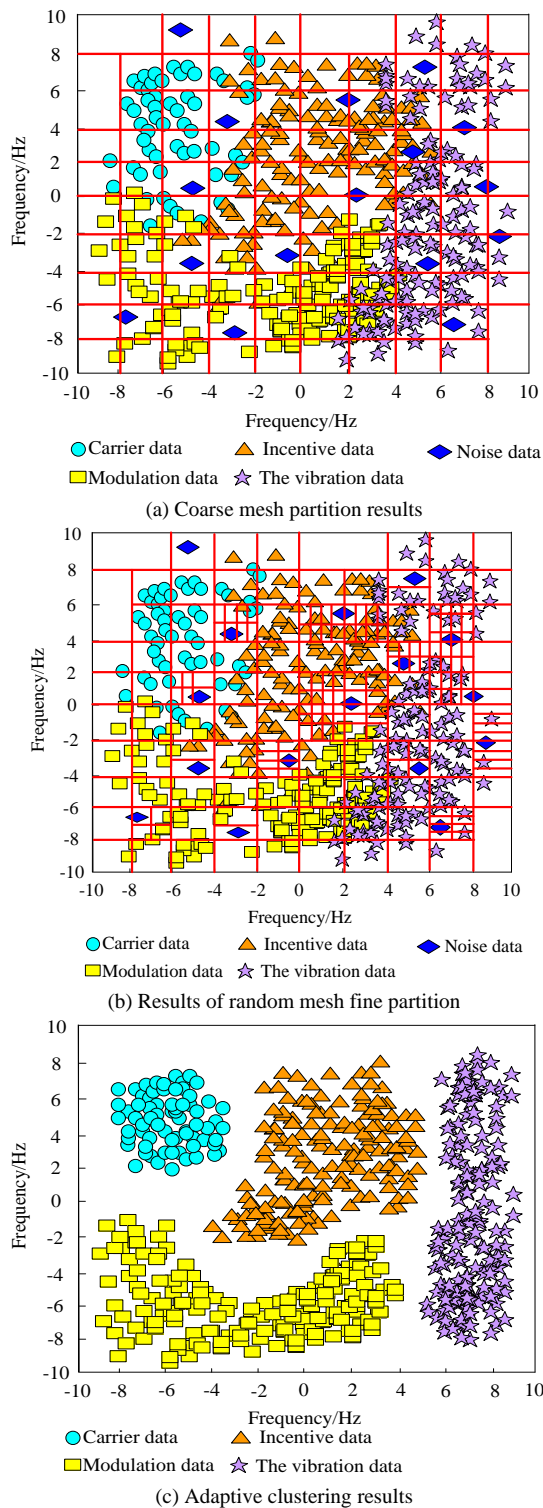


Fig. 4. Adaptive Partitioning Results of Project-based Learning Data in Low-dimensional Simulation Experiment Information Set.

The external evaluation indexes Rand index, Purity and normalized mutual information (NMI), internal evaluation indexes compactness (CP), separation (SP), and relative evaluation indexes including Davies bouldin index (DBI) and Dunn validity index (DVI) are used as evaluation indexes. The higher the value of Rand index is used to measure the

percentage of correct division, the higher the adaptive partitioning accuracy is, the value range is  $[0, 1]$ ; Purity represents the purity of the partitioning result, and the value range is  $[0\%, 100\%]$ . The higher the value is, the better the partitioning result is; NMI is to evaluate the consistency between the clustering results and the real categories through the mutual information between the clustering results and the real categories. The value is between  $[0, 1]$ . The higher the value is, the better the classification effect is; The lower CP means that the closer the clustering distance within the class is, the better the compactness is, and the value range is  $[0, 1]$ ; The higher the larger of SP is, the farther the cluster distance and the distance between clusters are, and the value range is  $[0, 1]$ ; The closer the intra class distance is, the farther the inter class distance is, the better the partitioning effect is; The smaller DBI means the smaller the intra class distance and the larger the inter class distance, the better the partitioning effect is, and the value range is  $[0, 1]$ ; The larger the DVI means the greater the distance between classes and the smaller the distance between classes, and the better the partitioning effect. The value range is  $[0, 1]$ ; The effect of adaptive partitioning of information set in simulation laboratory by the algorithm in this paper is analyzed. For high-dimensional information set of simulation experiment and low-dimensional information set of simulation experiment, the test results of adaptive partitioning of information set in simulation experiment by the algorithm in this paper are shown in Table I.

According to Table I, the RAND index of the algorithm in this paper is relatively high when adaptively dividing the information sets of low-dimensional and high-dimensional simulation laboratories, which is close to 1, indicating that the algorithm in this paper adaptively divides the information sets of different dimensions with high accuracy; The purity is also high, which is close to 100%, and the NMI is high, which is close to 1. This shows that the results of adaptive partitioning of different dimension information sets in this algorithm are very similar to the actual results; SP and DVI are both large, close to 1, CP and DBI are both small, and close to 0, which indicates that the algorithm in this paper adaptively partitions the information set with a large inter class distance and a small intra class distance, and has a better adaptive partitioning effect; Comprehensive analysis shows that for different dimension simulation laboratory information sets, the algorithm in this paper can accurately and adaptively partition the information sets, and has better adaptive partitioning effect of the information sets.

In order to further verify the performance of the algorithm studied in this paper, the algorithm in the literature [6] and the algorithm in the literature [7] introduced in the introduction are used as the comparison method to test the RAND index results in different dimensions, and the test is repeated five times. The results are shown in Table II.

It can be seen from Table II that under different dimensions, the lowest RAND index of this method is 0.96, close to 1. The RAND index of literature [6] method and literature [7] method is 0.89 and 0.88 respectively, which is far lower than that of this method. This shows that the adaptive division accuracy of this method is the highest and has certain applicability.

TABLE I. TEST RESULTS OF THE PROPOSED ALGORITHM FOR ADAPTIVE PARTITIONING OF SIMULATION EXPERIMENTAL INFORMATION SETS IN DIFFERENT DIMENSIONS

The evaluation index	Low-dimensional simulation laboratory information set	High dimensional simulation laboratory information set
Rand index	0.98	0.97
Purity	97.5%	98.2%
NMI	0.96	0.98
CP	0.01	0.02
SP	0.97	0.97
DBI	0.02	0.03
DVI	0.96	0.98

TABLE II. COMPARISON OF RAND INDEX RESULTS OF DIFFERENT METHODS IN DIFFERENT DIMENSIONS

group	Methods in this paper		Literature [6] Method		Literature [7] Method	
	Low-dimensional simulation laboratory information set	High dimensional simulation laboratory information set	Low-dimensional simulation laboratory information set	High dimensional simulation laboratory information set	Low-dimensional simulation laboratory information set	High dimensional simulation laboratory information set
1	0.98	0.97	0.89	0.87	0.88	0.81
2	0.99	0.96	0.88	0.88	0.81	0.75
3	0.97	0.98	0.85	0.84	0.86	0.78
4	0.98	0.97	0.86	0.85	0.72	0.79
5	0.99	0.97	0.89	0.82	0.83	0.82

#### IV. CONCLUSION

In order to deeply analyze the large amount of information generated by the simulation laboratory, it is necessary to accurately partition the information. An adaptive partition algorithm of simulation laboratory information set based on project-based learning data-driven and random grid is studied. The fuzzy C-means clustering algorithm in project-based learning data drive preliminarily divided the information, avoiding the constraints of time, space and other conditions. In the information space, the grid division algorithm of the data histogram was used to further divide the information, weakening the correlation between the data. Then, the grid division algorithm of the data histogram was used to cluster similar data, thus further enhancing the classification effect of the data. Through experiments, it is verified that the method can still maintain high classification accuracy in different dimensions, and has good application effect. However, there are still some shortcomings in this paper. Some heterogeneous data may appear in the data generated by the simulation laboratory, which is easy to be excluded as abnormal data when classifying. In future research, it is also necessary to strengthen the research on the classification and re clustering of heterogeneous data.

#### ACKNOWLEDGMENT

The study was supported by Shaanxi Province Education Science Planning Project --- "The research and practice of Project-based learning in higher education for English learning to improve the key competencies (Grant No. JYTYB2022-89)", and Shaanxi Higher Education Association Project --- "Research on education and teaching reform of non-

government undergraduate colleges under the background of new liberal arts" (Grant No.XGHZZ102).

#### REFERENCE

- [1] P. Sidjanin, J. Plavsic, I. Arsenic, and M. Krmar, "Virtual reality (vr) simulation of a nuclear physics laboratory exercise," *European Journal of Physics*, 2020, 41(6).
- [2] S. He, D. Kong, J. Yang, L. Ma, and Y.Chang, "Research on the teaching mode of university virtual laboratory based on component technology," *International Journal of Continuing Engineering Education and Life-Long Learning*, 2021, 31(1), 1.
- [3] M. D. Koretsky, "An interactive virtual laboratory addressing student difficulty in differentiating between chemical reaction kinetics and equilibrium," *Computer applications in engineering education*, 2020, 28(1), 105-116.
- [4] L. F. Zapata-Rivera, and C. Aranzazu-Suescun, "Enhanced virtual laboratory experience for wireless networks planning learning," *Revista Iberoamericana de Tecnologias del Aprendizaje*, 2020, PP(99), 1-1.
- [5] M. Ricaurte, and A. Vilorio, "Project-based learning as a strategy for multi-level training applied to undergraduate engineering students - sciencedirect," *Education for Chemical Engineers*, 2020, 33, 102-111.
- [6] A. Sutagundar, and P. Sangulagi, "Fog computing based information classification in sensor cloud- agent approach," *Expert Systems with Applications*, 2021, 182(2), 115232.
- [7] J. Flisar, and V. Podgorelec, "Improving short text classification using information from dbpedia ontology," *Fundamenta Informaticae*, 2020, 172(3), 261-297.
- [8] T. Zheng, Z. Duan, J. Wang, G. Lu, and Z. Yu, "Research on distance transform and neural network lidar information sampling classification-based semantic segmentation of 2d indoor room maps," *Sensors*, 2021, 21(4), 1365.
- [9] N. Berente, S. Seidel, and H. Safadi, "Data-driven computationally intensive theory development. *Information Systems Research*," 2019, 30(1), 50-64.

- [10] Y. L. Kang, L. L. Feng, and J. A. Zhang, "Cloud-Based Big Data Fuzzy Clustering Method Simulation Based on Grid Index," *Computer Simulation*, 2019, 36(12):341-344+441.
- [11] H. Liu, and Q. Qian, "Bi-level attention model with topic information for classification," *IEEE Access*, 2021, PP(99), 1-1.
- [12] Y. Song, L. Gao, X. Li, and W. Shen, "A novel point cloud encoding method based on local information for 3d classification and segmentation," *Sensors*, 2020, 20(9), 2501.
- [13] M. R. Bouadjenek, S. Sanner, and Y. Du, "Relevance- and interface-driven clustering for visual information retrieval," *Information Systems*, 2020, 94(6), 101592.
- [14] Z. Cai, X. Yang, T. Huang, and W. Zhu, "A new similarity combining reconstruction coefficient with pairwise distance for agglomerative clustering," *Information Sciences*, 2020, 508, 173-182.
- [15] M. B. Hajkacem, C. B. N'Cir, and N. Essoussi, "One-pass mapreduce-based clustering method for mixed large scale data," *Journal of Intelligent Information Systems*, 2019, 52(3), 619-636.
- [16] F. Cicalese, and E. S. Laber, "Information theoretical clustering is hard to approximate," *IEEE Transactions on Information Theory*, 2020, PP(99), 1-1.
- [17] X. Zhang, W. Pan, Z. Wu, J. Chen, and R. Wu, "Robust image segmentation using fuzzy c-means clustering with spatial information based on total generalized variation," *IEEE Access*, 2020, PP(99), 1-1.
- [18] C. Wu, and X. Zhang, "A novel kernelized total bregman divergence-driven possibilistic fuzzy clustering with multiple information constraints for image segmentation," *IEEE Transactions on Fuzzy Systems*, 2021, PP(99), 1-1.
- [19] D. Wei, Z. Wang, L. Si, C. Tan, and X. Lu, "An image segmentation method based on a modified local-information weighted intuitionistic fuzzy c-means clustering and gold-panning algorithm," *Engineering Applications of Artificial Intelligence*, 2021, 101(3), 104209.
- [20] M. S. Talib, A. Hassan, T. Alameri, Z. A. Abas, and N. Ibrahim, "A center-based stable evolving clustering algorithm with grid partitioning and extended mobility features for vanets," *IEEE Access*, 2020, PP(99), 1-1.

# Swine Flu Detection and Location using Machine Learning Techniques and GIS

P. Nagaraj<sup>1</sup>

Research Scholar of Osmania University and  
Associate Professor in Sreyas Institute of Engineering, and  
Technology, CSE  
Hyderabad, India

Dr. V. B. Narsimha<sup>3</sup>, Dr. B. Sujatha<sup>4</sup>

Assistant Professor  
Department of Computer Science & Engineering  
University College of Engineering (A), Osmania University,  
Hyderabad, India

Dr. A. V. Krishna Prasad<sup>2</sup>

Associate Professor in IT Department  
Maturi Venkata Subba Rao Engineering College  
Hyderabad, India

**Abstract**—The H1N1 virus, more commonly referred to as swine flu, is an illness that is extremely infectious and can in some cases be fatal. Because of this, the lives of many individuals have been taken. The disease can be transmitted from pigs to people. This research presents an artificial neural network (ANN) classifier for disease forecasting, as well as a technique for detecting people who are sick based on the geographic region in which they are found. The source codes for these two algorithms are provided below. These coordinates serve as the foundation for the GIS coordinates that are utilized in the method for assessing the extent to which the illness has spread. The ICMR and NCDC datasets were utilized in the study. They used Dynamic Boundary Location algorithm to detect swine flu affected person's location, the researchers discovered that the accuracy of the proposed classifier was 96 standard classifiers.

**Keywords**—Swine Flu; influenza; machine learning; GIS; classifiers; ANN; virus; algorithm

## I. INTRODUCTION

The swine flu is very contagious and spreads quickly. Potential vectors for the propagation of the disease include air and water. The influenza virus strain is also known as H1N1, which is another name for the H1N1 virus. Virus outbreaks might be global or local in scope, but they invariably end in the loss of human life. More than one country has reported human cases of influenza virus infection. It affects the respiratory system in humans. The WHO estimates that the influenza virus kills between 250,000 and 500,000 people each year, infecting 5 to 15% of the world's population and causing respiratory illnesses and other complications. Due to influenza-related infections, the United States of America loses between \$70 billion and \$170 billion. This has a huge impact on the economics of the rest of the globe [1] [2].

An artificial neural network (ANN)-based system for swine flu categorization and forecasting is suggested in this

paper. GIS coordinates of the person who is infected are acquired, and the second approach is used to locate the individual's position, as outlined in the article. The algorithm determines how far the sickness may spread, and it is constantly being updated as more people become ill. Confinement zones are defined by a set of boundaries generated by an algorithm. Machine learning may help in the analysis and exploration of patterns in a dataset. Machine learning-based classifiers take in data from training datasets and then use that data to make predictions.

Contribution from us: We present a cutting-edge technique that use machine learning to reliably forecast the incidence of swine flu in real time. Sore throats, chills, weariness, nausea, runny noses, body pains, coughs, and fevers are all common symptoms of swine flu infections. In certain cases, a cough may accompany these symptoms. ANN classifiers are able to predict whether or not a person has swine flu with a "Yes" or "No" answer, despite the difficulty of the task. We've come up with a supervised classifier as a solution to this problem. The classifier's output is sent into the proposed boundary algorithm, which generates the disease's sphere of influence's outer border.

## II. LITERATURE REVIEW

The Table I shows the Literature review of swine flu affected areas and the limitation of each paper. In this, we explain the main research gap of the latest significant state of the art approaches. Based on Table I we found research gaps in Machine learning for predicting location. We have seen various kinds of research papers and we conclude that most of the research papers are used Support vector machine, Random forest, Artificial Neural Network, NavieBayes, Ad boost and KNN. The main gap is quality of dataset is not used, accuracy and no comparisons are not made.



TABLE I. LITERATURE REVIEW

Ref No	Approach	Limitations	Gaps
[3]	Describe the use of social network plat forms to track people infected with swine flu based on their post on socialmedia like twitter	Social media data like twitter is not trustworthy	No detection method is discussed
[4]	Discusses the symptoms of the disease along with the hotspot detection	No comparisons made	No proper dataset And detection mechanism
[5]	Discusses 7 ML Classifiers for influenza detection	Dataset was randomly selected and consisted of 31268 records over a period from 2008-11	The number of records and time period both are less
[6]	Study and review of the Existing ML techniques that could detect other diseases were performed. The techniques like SVM, Random Forest, ANN, Naïve Bayes and Adaboost are used.	No classifier was proposed. Only standard base line algorithm were used	A quality dataset was not used for the study
[7]	Use deep learning neural network architecture for predicting thoracic disease using x-ray images. 50 layers Resnet architecture is used. Dynamic routing approach is used in between convolution layers	No classifier was Proposed. Data set is specific to application	Semi supervised approach not used in location information
[8]	Predicts and localization of disease done simultaneously. The use of deep learning classifier is used	Classifier is not explained	A quality dataset was not used for the study
[9]	Larger dataset are required to get better results	Gaussian distribution not used	No comparisons made
[10]	Images of pneumonia may be identified and pinpointed thanks to a model constructed using deep learning. Utilizes a segmentation-based approach.	Dataset was small	Average segmentation approach is used
[11]	It has been hypothesised that there exists a system that is capable of concurrently learning discriminative brain-region localization and sickness detection. The data came from ADNI, which was used as the source. The algorithm that will be used to categorise the data has been suggested.	The model works well for small patches	Classifier for prediction is not available
[12]	Proposed a classifier based on Random forest algorithm on individual symptoms of swine flu. Probability [reduction of each symptom was detected leading to the disease of swine flu	Dataset of GCI used is small	Did not perform proper clustering to detect hotspot of the disease
[13]	Create a concept for an artificial neural network (ANN) that loops back on itself. In comparative testing, the performance of the suggested method is Superior than that of SVM and Naïve Bayes.	Algorithm and dataset not described properly	No comparisons made
[14]	Discusses the methods like density estimation, Model based approaches to clustering	Detection process is not mentioned	A quality dataset was not used for the study
[15]	Uses Machine learning Techniques to find hotspot in fabrication technology. A combination of classification and feature extraction process are used	SVM kernel was used with an accuracy of 78%	Accuracy is low
[16]	Detects hotspot in online Forums. The proposed algorithm works in collaboration with K Means and SVM. The centre of the cluster is the hotspot	Apart from Detection. No proper algorithm is used to identify the hotspots	No comparisons made
[17]	Uses KNN algorithm to find hotspot with ML algorithm and	A proper dataset is not used	Segmentation of cluster region not specified
[18]	Applied a ML-based hotspot identification approach that includes lithography data into the development of the SM during the learning phase. True alarms are few and far between.	Did not detect all hotspots together	No comparisons made
[19]	Used feature extraction with tensor generation in CNN that has spatial relationships. The learning process is further extended to batch learning. Used gradient decent approach with low false alarm	Comparison Between different methods learning's not specified	A quality dataset was not used for the study

### III. METHODOLOGY

#### Dataset

The ICMR supplied the data utilized in the study. 122751 cases of swine flu have been verified and 115517 people have recovered from the infection, making up 1264443 entries in the database. Between 2015 and 2021, 7336 Americans lost their lives as a result of various causes around the country. Shown in Table II are the dataset's most essential features.

TABLE II. DATASET

No of Records	1264443
Positive Cases	122751
Recovered	115517
Deaths	137234
No of Training Records	1002200
No of Test Records	250551

The Table III shows the attributes that was selected for the study after performing a feature selection. 10 attributes were selected for the study.

TABLE III. ATTRIBUTES SELECTED FOR THE STUDY

Sl. No.	Symptoms/Attributes
1	Fever
2	Chills
3	Fatigue
4	Body/Muscle ache
5	Loss of Appetite
6	Headache
7	Dry Cough
8	Sore throat
9	Running/stuffy nose
10	Age

The Proposed System architecture of the system is displayed in Fig. 1. The ICMR and NCDC data sets supply the information that is used in the recommended technique. The recommended approach can determine whether or not a person is sick, and hence at danger, when GIS coordinates are used to find a likely hotspot for the Swine Flu. The purpose of this research is to identify cases of swine flu using the application of machine learning. This goal has been accomplished as a direct consequence of implementing the classifier that was recommended [30].

For the purpose of anticipating Swine Flu symptoms, this system makes use of an implementation that is based on artificial neural networks. Even though there are 10 separate tones being sent into the network in Fig. 2, it only produces two distinct tones. The missing layer, which is composed of two hidden levels, has had a total of eight nodes removed from it. The artificial neural network (ANN) is a type of classifier that sorts incoming data by employing a multilayer feed forward back propagation method. The technique of modeling

makes use of the activation function that is associated with the sigmoid function. Using this procedure, swine flu can be identified in one of two different ways: Training and tests are included in this package. Each individual neural network must be trained on all ten characteristics that comprise a record in order to function properly. Please be aware that this is an activity that will be logged.

The Table IV show the parameters of Artificial Neural Network(ANN), it takes 10 input nodes, two output nodes , two hidden layer( each hidden layer contain four nodes)and used backpropagation method. The Activation function is Sigmoid function.

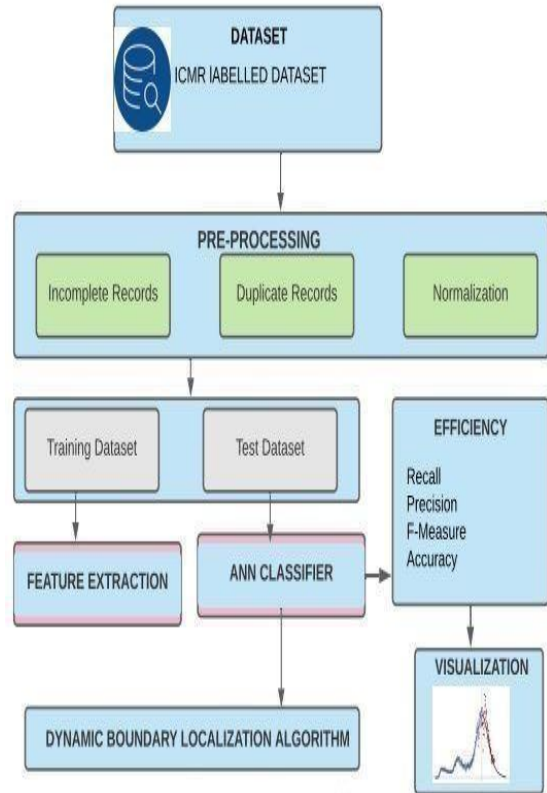


Fig. 1. Proposed System Architecture.

TABLE IV. PARAMETERS OF THE ANN

Number of input nodes	10
Number of output nodes	2
Number of hidden layers	2
Number of nodes in hidden layer	8
Paradigm	Multilayer Feed-Forward Network (Backpropagation)
Weight updating rule	Delta weight
Activation function	Sigmoid function

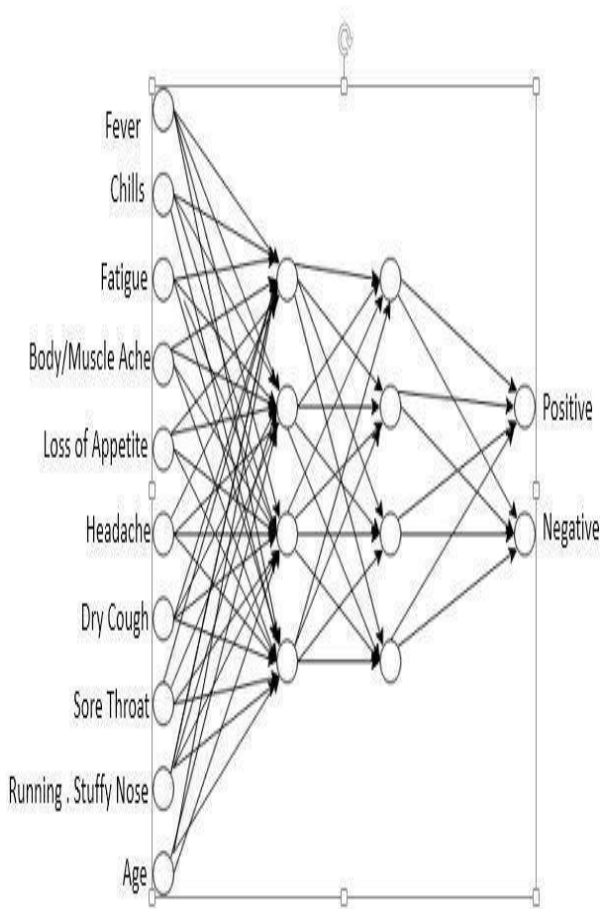


Fig. 2. Flow of ANN using Selected Features.

The Fig. 3 describes the how Artificial Neural Network(ANN) works.

If you go through the data set, you will learn about the classes. To categorise all records, this process is repeated as many times as necessary. Errors in training will be propagated backward and new penalties or delta error values will be applied if necessary. The figure shows how the algorithm for making predictions works in a visual way. Next, a random weight is supplied to each node in the artificial neural network to "initialize" it. Every node has been given a certain threshold. Calculating how much of a miscalculation occurred begins by determining the difference between projected and actual data. Efforts are taken to minimize the chance of making a mistake. An optimization condition must be met in order for checks to be performed. The output error can also be affected by errors that occur in the hidden layers. At every step of this procedure, it is necessary to use the gradient Descent function. It is important to meet the erroneous criteria in order to generate an accurate forecast for swine flu. A total of four data sets are available for training [31].

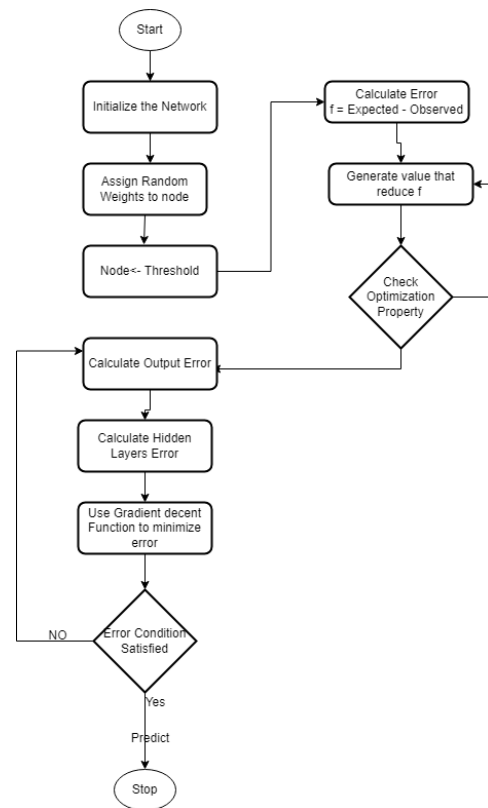


Fig. 3. Flowchart of the Proposed Algorithm.

### Proposed Predicting Algorithm

#### Algorithm Prediction Detecting Diseases Formation

Input Weights  $w_i$  node Threshold

L1 Begin

L2 {

L3 Initialize the network

L4  $W_i \leftarrow$  Random

L5  $Node \leftarrow$  Threshold

L6  $f = Output\ Expected - Output\ Observed$

L7  $f_1 =$  Set of  $f \leftarrow$  minimize error

L8 Repeat L6 to L7 until optimization criteria

L9 End

L10  $Output\ Error \leftarrow \frac{1}{n} \sum_{i=1}^n (Expected - Output)^2$

L11 Hidden error  $\leftarrow$  Back Propagation method

L12 If Error criteria checking  $\rightarrow$  False Go to Line 10

L13 Repeat Until Expected = Output

L14 }

L15 End

For the suggested classifier, a random seed is used to initialize the ANN network's weights. Each node has its unique threshold. The node is activated when the reset threshold is met. The total number of mistakes in each node's output is calculated. By using it, we may get an accurate estimate of the variance between what was expected and what was really achieved, which is known as the error value. Each

node in the network calculates the ANN's error. By utilizing the new value, we may achieve a reduced level of node-level inaccuracy. Errors should be kept to a minimum as the primary goal of the project. To determine hidden layer errors, the output error is first decreased.

The gradient descent function has been used to lower the total amount of error in our calculations. It is still possible for the gradient descent process to reach a point of convergence even while the data cannot be separated linearly. Each time an incorrect number is rectified, a little amount is deducted from the overall sum. When an error number rises, the error value that was there just before the rise is displayed on the screen as well. Because of this, the function is referred to as "decent." The gradient descent technique can be used to break out of a loop if an error continues to occur [32].

For starters, only 20% of the data is really used for training. Our second iteration uses the first 20% of data, as well as the remaining 20% of data. More than half of the records are still in use in the third iteration. We were able to use about 80% of the data in the final dataset. Approximately 20% of the entries in the database are taken from which a sample of the full database may be utilized for testing. At the conclusion of each cycle, the accuracy or detection rate is shown in the Table V. With a 96 percent accuracy rate, the classifier is clearly doing its job. For your perusal, Table 5 displays the matrix of perplexity. There were 39450 positive test findings, and 1861 27 negative test results were uncovered in this investigation. The values are shown in their proper

Training Data Accuracy Rate

TABLE V. SYSTEM ACCURACY RATE

Iteration	Upper value for -ve case	Lower value for +ve case	Required Threshold value	Number of Records per iteration	No. of correctly classified patterns	Accuracy for each partition
1	0.287656	0.565139	0.4263975	202310	192196	78.00%
2	0.3315276	0.524858	0.4281928	404621	192194	76.00%
3	0.7904	0.379523	0.584976	606932	212425	84.00%
4	0.878156	0.335547	0.6068515	809243	232656	96.00%

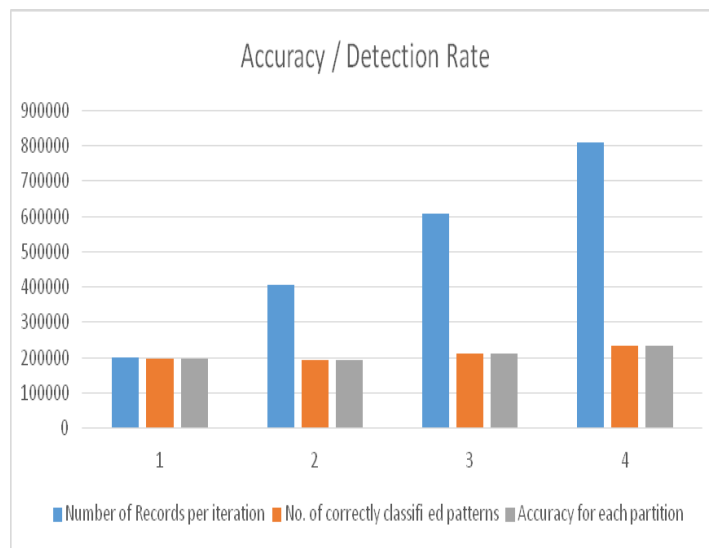


Fig. 4. Graph of Accuracy/Detection Rate.

places on the table. Classifier's findings from past study are summarized in the table that follows. 96 percent of the time, the proposed neural network-based classifier works as expected. The Fig. 4 shows the accuracy and detection graph [33].

Testing data Confusion Matrix:

It is standard practice to use Table VI a confusion matrix to explain the performance of a classification model (also known as a "classifier") on a set of test data for which the actual values are known and finally Table VII shows the classifier values accuracy is 0.96[34].

Classifier Values:

TP: True Positive

TN: True Negative

FP: False Positive

FN: False Negative

Performance of Training and Testing

The Fig. 6 shows the performance of training and testing i.e., Training loss and epochs,

The Accuracy rate of Training and Testing, here we execute code in python programming the Epoch are from one to five. Here the final loss 0.1111 and accuracy rate is 96% in Fig. 5.

TABLE VI. CONFUSION MATRIX

Correct P	Incorrect N	Result
39450	9217	positive
10184	186127	negative

TABLE VII. CLASSIFIER VALUES

Recall	TP/(TP+TN)	0.17488
Precision	TP/(TP+FP)	0.79481
F-measure	(2*Precision*Recall)/(Precision+Recall)	0.448440
Accuracy	(TP+TN)/(TP+FN+FP+TN)	0.96690

```

Epoch 1/5
600/600 [=====] - 17s 27ms/step - loss: 1.5885 - accuracy: 0.5643
Epoch 2/5
600/600 [=====] - 16s 27ms/step - loss: 0.2812 - accuracy: 0.9184
Epoch 3/5
600/600 [=====] - 17s 28ms/step - loss: 0.1982 - accuracy: 0.9422
Epoch 4/5
600/600 [=====] - 17s 28ms/step - loss: 0.1567 - accuracy: 0.9546
Epoch 5/5
600/600 [=====] - 17s 28ms/step - loss: 0.1288 - accuracy: 0.9629
313/313 [=====] - 1s 2ms/step - loss: 0.1111 - accuracy: 0.9669
    
```

Fig. 5. Results of Training and Testing.

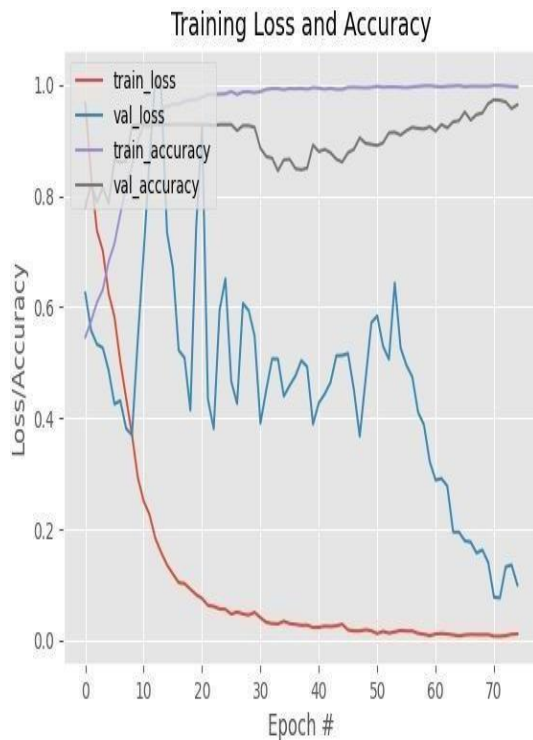


Fig. 6. Performance of Training and Testing.

#### IV. PROPOSED DYNAMIC BOUNDARY LOCATION ALGORITHM

Pseudo code

Algorithm Dynamic Boundary Location Algorithm INPUT

ICMR Dataset, Distance D=50, GIS Co-ordinates

L1 Begin

L2 Read-> Training Data, Distance, GIS Co-ordinates L3 For

L4 Each training data record L5 Do

L6 {

L7 Mark first data point ->X

L8 Identify points B1, B2, and B3 -> D/2 from X L9 Plot

point ->GIS

L10}

L11 End Do

L12 if region of B1 B2 B3 locate ->X1 L13 X1<- Shortest

Distance point-> X L14 Else

L15 X1<- Outside region B1 B2 B3

L16 Locate nearest Boundary point <- X1 L17 Repeat L5 to

L10

L18 until END of training data record

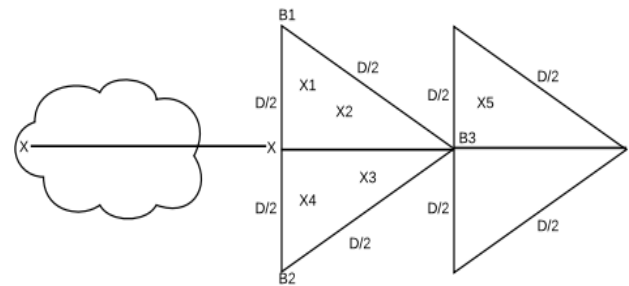


Fig. 7. Working of Proposed Dynamic Boundary Localization Algorithm.

This Fig. 7 article provides an explanation of the Dynamic Boundary Localization technique. To phrase it in a different way: Imagine that X is the location where the first case of the flu or swine flu was found. As a point of reference, Google Maps is utilized, and the value of the country/region/Distance area is selected based on the location of the point X. Let's say D represents a distance of 50 kilometers in this scenario. From point X, mark the locations of points B1, B2, and B3 at a distance of D/2 in the north, south, and east directions, respectively. This location houses the outermost cluster. The maximal possible spread of the illness. Find the data point that is geographically closest to each point (x1, x2, etc.) that makes up the triangulated region so that you may create tiny clusters or joint clusters. If a data point is located outside of the region that was triangulated, this indicates that the disease has spread to a location that is not included in the triangulation. In this stage, the algorithm will need to dynamically expand the area. This is accomplished by looking for the boundary point that is geographically closest to the data point that is located outside the boundary. The boundary points X3 and B3 are the ones that are located the most closely to one another. It is necessary to repeat the stages from B3 to B4 to B5 to B6 once more. The algorithm first creates a zone in the shape of a

triangle, and then it expands to the north, south, and east sides of the triangle. If it grows on just three of the globes or Earth's sides, then it will encompass all of the other four sides [35].

V. RESULTS AND DISCUSSION

The suggested classifier has a consistent upward trend over the length of the learning process. According to Table IV, there was a success rate of 78% in the first 20% of data that were looked at. The right classification was applied to a total of 197252 items. The accuracy of the classifier increases from 80 percent after the third iteration to 84 percent after the fourth iteration, and then to 92 percent after the fifth iteration. We can ensure that our classifiers continue to improve as we increase the number of iterations that we feed into them as well as the quantity of training data that we feed into them by utilizing the techniques of machine learning [36]. This is an easy assignment to do because the ANN classifier that was explained has an accuracy rate of 96 percent. Fig. 4 provides a visual representation of the ANN (artificial neural network) prediction classifier detection rate. Table V presents the matrix of ambiguity for your review and consideration. In the database, there are a total of 39450 positive entries in addition to 186127 negative records [37].

The Table VIII show the comparison of results with existing literature i.e., it will be compared with Machine learning algorithms and get the accuracy results. The Dynamic Boundary Location Algorithm (DBLA) and finally we get the 96 accuracy.

TABLE VIII. COMPARISON OF RESULTS WITH EXISTING LITERATURE

Literature	Approach	Accuracy %
[12] Kakulapati et al. 2020	ANN and Random Forest algorithm	87
<b>Proposed D B L A classification method</b>		<b>96</b>
[8] Li et al. 2013	Bayesian and Markov network	83
[20] Xue et al. 2018	Regression and ANN	81
[21] Biswas et al. 2015	ANN classifier	78
[22] Srinivas et al. 2018	Naive Bayesian Classifier	84
[3] Kostkova et al. 2014	Cross correlation	76
[23] Volkova et al. 2017	Neural networks	87
[24] Raval et al. 2016	feed-forward neural network construction	73
[25] Singh and Kaur et al	support-vector-regression	91
[26] Tate et al. 2017	Random forest	78
[27] Byrd et al. 2016	Web based sentimental analysis Twitter	86
[28] Xue et al. 2019	support-vector-regression	89
[29] Rao et al. 2021	Hybrid voting algorithm	73

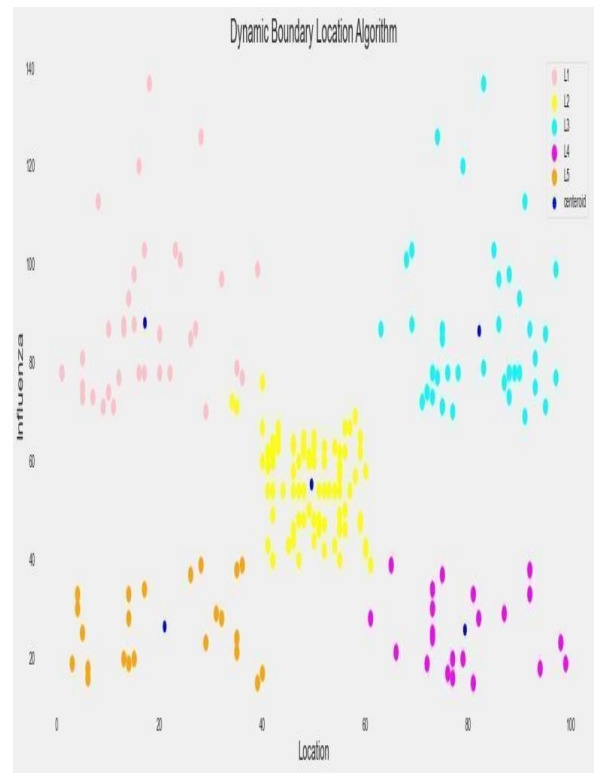


Fig. 8. Results of the Second Technique was Provided for Dynamic Border Localization.

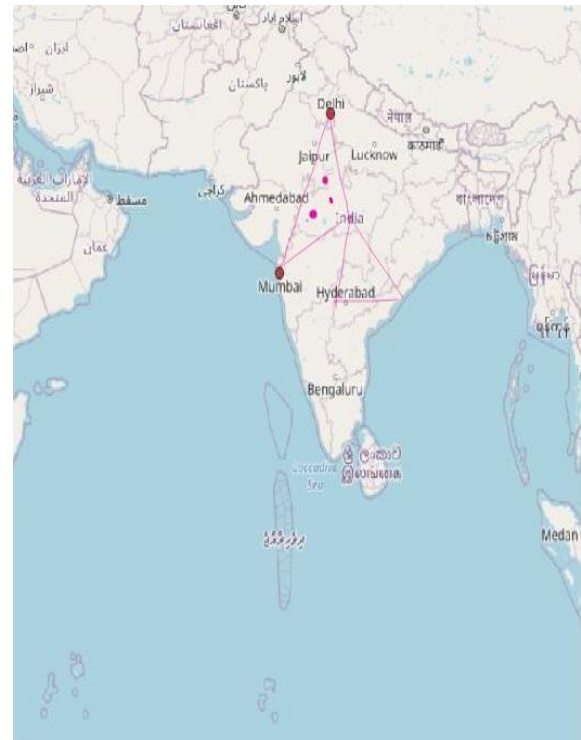


Fig. 9. The Output from Both the Existing Standard Literature and the GIS Tools.



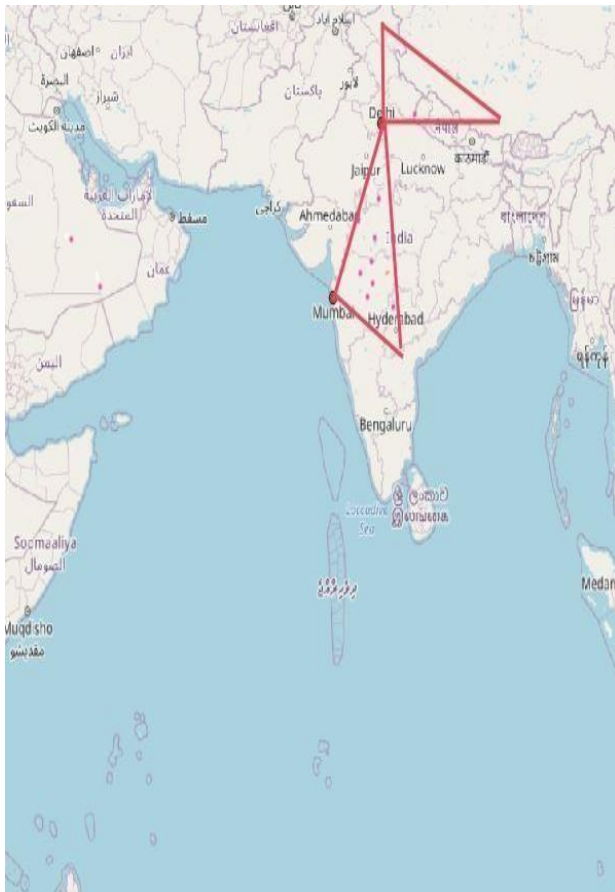


Fig. 10. Localization of Confirmed Swine Flu Cases using Proposed Dynamic Boundary Location Algorithm.

The total number of genuine positives was 9217, whereas the number of erroneous negatives was 10184. The method for predicting the spread of swine flu has an accuracy rate of 96 percent overall. According to the findings, the accuracy of the classifier is measured at 17.4 percent recall. According to the findings of the test, the accuracy was found to be 79.48 percent, and the F-measure was found to be 44.84 percent. The results of the experiments are presented in Table VII, which also serves as a comparison of the experimental outcomes of the recommended classifier with previously published research. The ANN classification algorithm that has been presented in this article cannot be compared in any way to the approaches that have been discussed in the previous paragraphs. Fig. 8 illustrates the results of the second technique that was provided for dynamic border localization. This method was proposed before [38]. The Fig. 9 and Fig. 10 show the output from both the existing standard literature and the GIS tools that are now available thanks to the method's superior area localization accuracy and GIS co-ordinate precision.

## VI. CONCLUSIONS

The use of localization for predictive purposes about swine flu is covered in great detail in the study. Both prediction and location may be accomplished with the use of algorithms. A back propagation classifier based on an artificial neural network should be considered as a first step. The Indian

Council of Medical Research (ICMR) is the source of the data that feeds the algorithm. The classification process is broken up into two distinct steps by the classifiers. In the first step, instances of swine flu are identified, and in the second stage, positive cases are localized with the use of an algorithm known as the dynamic boundary. The locations in the world where this extremely contagious virus is spreading may be easily identified, and containment zones may be set up in regions where the prevalence of the disease is high. Eighty percent of the data in the dataset is utilized in the training of the classifier. During the testing phase, there are an infinite number of potential combinations and permutations that might result in a record testing positive or negative. In order to assess the level of accuracy achieved by the methodology, a wide variety of well-established machine learning approaches were used as benchmarks. The recommended classifier has a detection rate that was significantly higher than the detection rates of the typical classifiers that are currently being utilized. The likelihood that the algorithm has gotten the answer correct is 96 percent. In order to evaluate the precision of the dynamic boundary method, many different GIS tools were utilized. It has been demonstrated that the localization accuracy achieved using the dynamic boundary technique is on par with that achieved using more conventional GIS tools.

## REFERENCE

- [1] J. Li and C. Cardie, "Early Stage Influenza Detection from Twitter," *arXiv:1309.7340 [cs]*, Nov. 2013, Accessed: Nov. 25, 2021. [Online]. Available: <http://arxiv.org/abs/1309.7340>.
- [2] R. Singh and R. Singh, "Applications of sentiment analysis and machine learning techniques in disease outbreak prediction – A review," *Materials Today: Proceedings*, p. S2214785321032764, May 2021, doi: 10.1016/j.matpr.2021.04.356.
- [3] P. Kostkova, M. Szomszor, and C. St. Louis, "#swineflu: The Use of Twitter as an Early Warning and Risk Communication Tool in the 2009 Swine Flu Pandemic," *ACM Trans. Manage. Inf. Syst.*, vol. 5, no. 2, pp. 1–25, Jul. 2014, doi: 10.1145/2597892.
- [4] R. Chen, W. Zhong, H. Yang, H. Geng, X. Zeng, and B. Yu, "Faster Region-based Hotspot Detection," in *Proceedings of the 56<sup>th</sup> Annual Design Automation Conference 2019, Las Vegas NV USA, Jun. 2019*, pp. 1–6. doi: 10.1145/3316781.3317824.
- [5] A. López Pineda, Y. Ye, S. Visweswaran, G. F. Cooper, M. M. Wagner, and F. (Rich) Tsui, "Comparison of machine learning classifiers for influenza detection from emergency department free-text reports," *Journal of Biomedical Informatics*, vol. 58, pp. 60–69, Dec. 2015, doi: 10.1016/j.jbi.2015.08.019.
- [6] Laura K. Borkenhagen, Martin W. Allen & Jonathan A. Runstadler (2021) Influenza virus genotype to phenotype predictions through machine learning: a systematic review, *Emerging Microbes & Infections*, 10:1, 1896-1907, DOI: 10.1080/22221751.2021.1978824.
- [7] Y. Shen and M. Gao, "Dynamic Routing on Deep Neural Network for Thoracic Disease Classification and Sensitive Area Localization," *arXiv:1808.05744 [cs]*, Aug. 2018, Accessed: Dec.07,2021.[Online]. Available: <http://arxiv.org/abs/1808.05744>.
- [8] Z. Li et al., "Thoracic Disease Identification and Localization With Limited Supervision," p. 10. 2013.
- [9] W. Huang, C. Yang, and T. Hou, "Spine Landmark Localization with combining of Heatmap Regression and Direct Coordinate Regression," p. 6.
- [10] R. Amer, M. Frid-Adar, O. Gozes, J. Nassar, and H. Greenspan, "COVID-19 in CXR: from Detection and Severity Scoring to Patient Disease Monitoring," *IEEE J. Biomed. Health Inform.*, vol. 25, no. 6, pp. 1892–1903, Jun. 2021, doi: 10.1109/JBHI.2021.3069169.
- [11] C. Park and H.-I. Suk, "Deep Joint Learning of Pathological Region Localization and Alzheimer's Disease Diagnosis," *arXiv:2108.04555 [cs]*, Aug. 2021, Accessed: Dec. 07, 2021. [Online]. Available:

- <http://arxiv.org/abs/2108.04555>.
- [12] V. Kakulapati, V. K. Kumar, V. S. Srikar, Y. S. Rao, and T. O. Edoh, "An intelligent framework of Swine flu status prediction by random forest algorithm," p. 8.
- [13] D. Bhatt, D. Vyas, M. Kumhar, and A. Patel, "Swine Flu Prediction Using Machine Learning," in *Information and Communication Technology for Intelligent Systems*, vol. 107, S. C. Satapathy and A. Joshi, Eds. Singapore: Springer Singapore, 2019, pp. 611–617. doi: 10.1007/978-981-13-1747-7\_60.
- [14] A. B. Lawson, "Hotspot detection and clustering: ways and means," *Environ Ecol Stat*, vol. 17, no. 2, pp. 231–245, Jun. 2010, doi: 10.1007/s10651-010-0142-z.
- [15] Y.-T. Yu, G.-H. Lin, I. H.-R. Jiang, and C. Chiang, "Machine-Learning-Based Hotspot Detection Using Topological Classification and Critical Feature Extraction," *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.*, vol. 34, no. 3, pp. 460–470, Mar. 2015, doi: 10.1109/TCAD.2014.2387858.
- [16] N. Li and D. D. Wu, "Using text mining and sentiment analysis for online forums hotspot detection and forecast," *Decision Support Systems*, vol. 48, no. 2, pp. 354–368, Jan. 2010, doi: 10.1016/j.dss.2009.09.003.
- [17] M. U. Ali et al., "Early hotspot detection in photovoltaic modules using color image descriptors: An infrared thermography study," *Int J Energy Res*, p. er.7201, Aug. 2021, doi: 10.1002/er.7201.
- [18] J. W. Park, A. Torres, and X. Song, "Litho-Aware Machine Learning for Hotspot Detection," *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.*, vol. 37, no. 7, pp. 1510–1514, Jul. 2018, doi: 10.1109/TCAD.2017.2750068.
- [19] H. Yang, J. Su, Y. Zou, Y. Ma, B. Yu, and E. F. Y. Young, "Layout Hotspot Detection With Feature Tensor Generation and Deep Biased Learning," *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.*, vol. 38, no. 6, pp. 1175–1187, Jun. 2019, doi: 10.1109/TCAD.2018.2837078.
- [20] H. Xue, Y. Bai, H. Hu, and H. Liang, "Influenza Activity Surveillance Based on Multiple Regression Model and Artificial Neural Network," *IEEE Access*, vol. 6, pp. 563–575, 2018, doi: 10.1109/ACCESS.2017.2771798.
- [21] S. Kr. Biswas, B. Baruah, B. Purkayastha, and M. Chakraborty, "An ANN based Classification Algorithm for Swine Flu Diagnosis," *IJKBCS*, vol. 3, no. 1, 2015, doi: 10.21863/ijkbcs/2015.3.1.005.
- [22] P. Srinivas, D. Bhattacharyya, and D. M. Chakkaravarthy, "An Artificial Intelligent based System for Efficient Swine Flu Prediction using Naive Bayesian Classifier," *IJCRR*, vol. 12, no. 15, pp. 134–139, 2020, doi: 10.31782/IJCRR.2020.121519.
- [23] S. Volkova, E. Ayton, K. Porterfield, and C. D. Corley, "Forecasting influenza-like illness dynamics for military populations using neural networks and social media," *PLoS ONE*, vol. 12, no. 12, p. e0188941, Dec. 2017, doi: 10.1371/journal.pone.0188941.
- [24] D. Raval, D. Bhatt, M. K. Kumhar, V. Parikh, and D. Vyas, "Medical Diagnosis System Using Machine Learning," vol. 7, no. 1, p. 6, 2015.
- [25] S. Singh and H. Kaur, "Influenza prediction from social media texts using machine learning," *J. Phys.: Conf. Ser.*, vol. 1950, no. 1, p. 012018, Aug. 2021, doi: 10.1088/1742-6596/1950/1/012018.
- [26] A. Tate, U. Gavhane, J. Pawar, B. Rajpurohit, and G. B. Deshmukh, "Prediction of Dengue, Diabetes and Swine Flu Using Random Forest Classification Algorithm," vol. 04, no. 06, p. 7.
- [27] K. Byrd, A. Mansurov, and O. Baysal, "Mining Twitter data for influenza detection and surveillance," in *Proceedings of the International Workshop on Software Engineering in Healthcare Systems*, Austin Texas, May 2016, pp. 43–49. doi: 10.1145/2897683.2897693.
- [28] H. Xue, Y. Bai, H. Hu, and H. Liang, "Regional level influenza study based on Twitter and machine learning method," *PLoS ONE*, vol. 14, no. 4, p. e0215600, Apr. 2019, doi: 10.1371/journal.pone.0215600.
- [29] N. T. Rao, D. Bhattacharyya, E. S. N. Joshua, and C. V. Satyanarayana, "Prediction of Swine Flu using a Hybrid Voting Algorithm," p. 9, 2021.
- [30] P. NAGARAJ, Dr. A. V. Krishna Prasad, "Survey on Swine flu Prediction," *International Journal of Management, Technology and Engineering*, Volume IX, Issue V, May/2019, ISSN No: 2249-7455, Page no: 937-941.
- [31] P. Nagaraj, Rajesh Banala and A. V. Krishna Prasad, "Real Time Face Recognition using Effective Supervised Machine Learning Algorithms," *Journal of Physics: Conference Series* 1998 (2021) 012007 IOP Publishing doi: 10.1088/1742-6596/1998/1/012007.
- [32] P. Nagaraj and Dr. A. V. Krishna Prasad, "A Novel Technique to Detect the Hotspots Swine Flu Affected Regions", Published in: 2021 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), 15 November 2021 DOI: 10.1109/ICRITO51393.2021.9596422, Electronic ISBN: 978-1-6654-1703-7 CD: 978-1-6654-1702-0.
- [33] Nagaraj P, Dr. A. V. Krishna Prasad, "A Cloud Computing Emerging Security Threats and Its Novel Trends in Knowledge Management Perception", *International Journal of Emerging Technology and Advanced Engineering Website: www.ijetae.com* (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 7, Special Issue 2, December 2017).
- [34] Nagaraj P, Rohit Kumar K, Rajesh Banala b, "Energy efficient 2 tier data aggregation scheme in Sensor networks", Accepted 7 March 2021, <https://doi.org/10.1016/j.matpr.2021.03.1402214-7853/> 2021 Elsevier Ltd. scientific committee of the Emerging Trends in Materials Science, Technology and Engineering.
- [35] P. Nagaraj, Gunta Sherly Phebe, Anupam Singh, "A Novel Technique to Classify Face Mask for Human Safety", 2021 Sixth ICIP Published in: 2021 Sixth International Conference on Image Information Processing (ICIIP), 26-28 Nov. 2021, 10 February 2022. DOI: 10.1109/ICIIP53038.2021.9702607 Publisher: IEEE Conference Location: Shimla, India.
- [36] P. Nagaraj Rajesh Banala b, Vicky Nair "Performance of Secure Data Deduplication Framework in Cloud" (ICAIDS-2022) organized by Sreyas Institute of Engineering and Technology, Hyderabad, Telangana, India during March 11-12, 2022.
- [37] P. Nagaraj, Chenna Sriya, Sarikonda Mahidhar Raju, Saba Naazneen Kauser, Chintala Rakesh "SMART Street – An Artificial Intelligence (AI) Powered Street Garbage Detection & Alert System", *Design Engineering* ISSN: 0011-9342 | Year 2021, Issue: 9, Pages: 18131-18141.
- [38] P. Nagaraj, P. Sahith Krishna, P. Shiva Sai "Forecasting Cyber Attacks Using Machine Learning" *Journal of Opto Electronics Laser*, ISSN: 1005-0086, DOI: 10050086.2002.07.66, Volume 41, Issue 7, 2022, Pages: 550-556.

# Taxation Transformation under the Influence of Industry 4.0

Pavel Victorovich StroeV<sup>1</sup>, Rafael Valiakhmetovich Fattakhov<sup>2</sup>, Olga Vladimirovna Pivovarov<sup>3</sup>, Sergey Leonidovich Orlov<sup>4</sup>, Alena Stanislavovna Advokatova<sup>5</sup>

Financial University under the Government of the Russian Federation, Moscow, Russia<sup>1, 2, 3, 4, 5</sup>

**Abstract**—Today the growing level of automation and the new concept of online technologies are transforming the traditional industry. Value is generated with the help of Industry 4.0 technologies that not only increase the efficiency and agility of supply chains, create new products and offer new ways of connecting businesses and consumers but also have a major impact on the traditional tax system. This study aims at determining changes in the modern industrial economy and substantiating possible directions for transforming the tax system to adapt it to the requirements of Industry 4.0. The objective is to identify the relationship between the digitalization of the economy, the use of blockchain technologies, robotics, automation, M2M technologies offered by Industry 4.0, and taxation. The article demonstrates how these technologies influence taxes and proposes measures to address possible tax issues. The authors of the article have concluded that the reasons (and goals) for transforming the current tax system as a result of the development of Industry 4.0 technologies are as follows: 1) to increase or stabilize tax revenues to compensate for tax losses and finance new education needs; 2) to introduce innovations for the development of Industry 4.0 and further digitalization of the economy; 3) to create an automatic tax administration system.

**Keywords**—Digitalization; cryptocurrency; blockchain; robotics; automation; innovation; tax

## I. INTRODUCTION

New technologies, principles, and approaches to production, automation, and robotics are influencing the level of employment [1], the amount of operating costs [2, 3], tax revenues [4], and the traditional taxation system as a whole [5].

The existing taxation system is built on the principles that have not changed over a long time, namely: fairness (all people pay taxes in proportion to their income and capabilities); certainty (the elements and overall organization of tax payments are clear); convenience (the mechanism for paying taxes should be as easy as possible for taxpayers); efficiency (the cost of administering taxes should be minimum) [6].

In our opinion, technological progress and globalization determine other priorities for socio-economic development, the specifics of production and sources of income [7], transforming the structure of tax systems, the organization of tax collection, shifting emphasis from one taxation object to another, expanding tax bases or reducing benefits [8, 9]. Thus, we want to answer the following questions: what are the consequences of the modern taxation system for introducing

Industry 4.0 and the digitalization of the economy as a whole? How will the introduction of Industry 4.0 affect the taxation system's organization? How will Industry 4.0 technologies affect the tax structure and will new taxes be introduced?

## Literature Overview

Different scholars are trying to answer these questions about the impact of the changes brought by Industry 4.0 on the development indicators of enterprises, households, and countries as a whole. Within the framework of studies [10, 11], an attempt was made to determine the potential tax implications for enterprises using new technologies and approaches to their production. The impact of Industry 4.0 on the tax strategy is analyzed in [12, 13]. Some studies conducted by specialists in taxation [14, 15] dwell on tax evasion in the digital economy. Recently, more and more scientific works have been prepared on the relationship between new technologies (such as robotics and the blockchain system) and taxes [15, 16]. Much attention is paid to electronic tax administration and control [17, 18].

Since digitalization, robotics, M2M technologies, and blockchain lead to significant changes both in national tax systems and in international taxation [19], some scholars offer several approaches to solving emerging problems: 1) to apply taxation to new technologies and their products and applications, for example, to extend traditional taxes to such objects as personal data, cryptocurrencies, and imputed income of robots [20]; 2) to replace digital transactions and shortfalls in revenues by traditional taxation objects in the form of tangible assets and/or increase tax pressure and the degree of progressive taxes already levied on such objects [21]; 3) to build a new tax system and transfer it to automatic taxation using blockchain technologies [22].

We believe that participants in global economic processes are entering a new technological era, therefore the development of Industry 4.0 does not affect individual spheres but rather concerns the whole world community.

Thus, the article aims at studying changes in the modern industrial economy and substantiating possible directions for transforming the tax system for its adaptation in the context of the development of Industry 4.0.

The research tasks are as follows:

- To reveal the main technological changes as a result of the development of Industry 4.0 and the digitalization of the economy as a whole;

- To identify economic changes in connection with the development of Industry 4.0;
- To analyze the consequences for the tax sphere, the impact on taxation objects and tax bases, which will cause economic changes in connection with the development of Industry 4.0;
- To propose measures in the tax sphere that need to be taken to solve the problems that arise.

The research hypothesis is as follows: economic changes as a result of the development of Industry 4.0 technologies affect (and will affect) the tax policy and tax system, which requires the transformation of the modern tax system and the use of both traditional and innovative tax tools.

## II. METHODS

To solve the above-mentioned tasks, we used theoretical and empirical methods of research, tested in scientific research devoted to the economic and information spheres of social development. Among the theoretical methods, we include the collection and analysis of scientific sources on the topic [23]. We used empirical research methods (expert survey) to collect quantitative data [24] with mathematical processing of results using the Kendall concordance coefficient (W) [25].

The study was conducted in three stages from February to April 2022 at the Financial University under the Government of the Russian Federation.

At the first stage of the study, we examined scientific and analytical works on the research topic.

The analysis of the relevant publications allowed us to identify the main technological changes as a result of the development of Industry 4.0: the digitalization of the economy, the use of blockchain technologies, robotics, automation, M2M technologies (machine-to-machine, data transfer directly between devices).

At the second stage of the study, we communicated online with the experts. The expert survey was carried out in Russian via e-mail.

The e-mails containing the above-mentioned questions were sent to 68 respondents, including 39 employees of high-tech companies from the top 15 of the Techuspekh 2020, 14 employees of analytical and information technology departments of the central office of the Federal Taxation Service of Russia, and 15 lecturers of the Financial University under the Government of the Russian Federation. The respondents were asked to justify their answers in a free form. As a result, we received answers from 61 experts.

In connection with the research topic, the experts were asked the following questions:

- 1) What are the consequences of digitalization, the use of blockchain technologies, robotics, automation, and M2M technologies for the economy and the tax sphere?
- 2) What measures should be taken in the field of taxation to offset the negative (positive) consequences of this influence?

All the respondents were informed about the purpose of the survey and that the authors planned to publish its results in a generalized form.

When receiving the answers, we asked the experts, depending on the significance of emerging problems, to arrange the consequences for the economy and the tax sphere on a scale of order, and to assign points. After that, each consequence of the economy and the tax sphere was ranked according to the points assigned by the experts.

For a more objective analysis of the data obtained during the expert survey, the consistency of expert opinions was measured through the mathematical processing of the results using the Kendall coefficient of concordance (W):

$$W = 12S/n^2(m^3-m),$$

where S is the sum of the squared deviations of all the ranks of each consequence for the economy and the tax sphere from the average value; n is the number of experts; m is the number of estimated economic/tax consequences.

Then the information obtained during the expert survey was processed to determine the impact of each consequence on the economy and the tax sphere, as well as to build a rank transformation matrix and calculate the arithmetic average of impacts for each consequence of the development of Industry 4.0 for the economy and the tax sphere, respectively.

The final impacts determine the significance of the consequences of Industry 4.0 for the economy and the tax sphere from the viewpoint of experts. All calculations were carried out using Excel 365 programs and the Stattech online service (<https://stattech.ru/>).

## III. RESULTS

The analysis of the expert survey has revealed the main consequences in the economy and taxation due to the introduction of new technologies and forms of doing business (Table I).

TABLE I. THE KEY ECONOMIC AND TAX CONSEQUENCES OF INDUSTRY 4.0

Changes	Consequences					
	Economy	Ranking	Impact	Taxation	Ranking	Impact
Digitalization	Increasing the purchase and sale of digital services and digitized goods	1	0.27	Reducing tax revenues on the consumption of traditional goods and services	1	0.36
	Protecting personal and corporate data	3	0.14	Protecting personal and corporate data	4	0.14
	Growing transnational stateless income	4	0.11	Erasing the tax base when taxing profits upon concluding agreements with citizens of other countries without their physical presence in these countries	3	0.16

Use of blockchain technology	Transparent operations	6	0.05	Possibility of a fundamental change in the tax administration system based on the automatic calculation of tax liabilities and their withdrawal from accounts	5	0.02
	Free access to transaction information	8	0.02			
	Minimizing the risk of losing documents	9	0.03			
Robotics, automation, M2M	Reducing the number of low-skill jobs	7	0.04	The need to compensate for the losses of social taxes that are currently paid by those employed in production	2	0.32
	Lack of personnel	5	0.10			
	Growing unemployment and income inequality	2	0.24			

Note: based on the results of an expert survey

According to the calculation of the Kendall's coefficient of concordance (W) ( $W = 0.69$ ), it can be argued that the expert opinions coincide since the value of  $W > 0.5$  indicates the objectivity of the survey results. This circumstance allows determining the impact of economic and tax consequences in connection with the development of Industry 4.0.

In conformity with the calculation results, digitalization has the greatest impact on the economy and taxation. Thus, the most important consequences are an increase in the purchase and sale of digital services and digitized goods (0.27), and the associated decrease in tax revenues on the consumption of traditional goods and services (0.36).

The robotics and automation of production which increase unemployment and income inequality (0.24) are no less important. As a result, it is necessary to compensate for losses in social taxes that are paid by those employed in production (0.32).

TABLE II. THE MEASURES TO BE TAKEN TO SOLVE TAX PROBLEMS

Changes	Consequences for the tax sphere	Measures to be taken to solve problems
Digitalization	Reducing the amount of tax revenues on the consumption of traditional goods and services	Introduction of a tax on digital goods and services or expansion of the existing tax base
	Protection of personal and corporate data	Introduction of a tax on the collection and use of personal data for Big Data owners
	Erosion of the tax base when taxing profits upon concluding agreements with citizens of other countries without their physical presence in these countries	Alignment of national legal norms with international tax legislation by improving transfer pricing for digital goods and services
Use of blockchain technologies	Possibility of a fundamental change in the tax administration system based on the automatic calculation of tax liabilities and their withdrawal from accounts	Development of a plan, tools and methods for implementing blockchain technologies for automated tax collection and unification of tax administration
Robotics, automation, M2M	Need to compensate for the losses of social taxes that are currently paid by those employed in production	Determining the possibilities of introducing new compensatory forms of taxes (tax on robots, universal basic dividend, etc.). Introduction of a tax credit for education and retraining loans

Note: based on the results of an expert survey

Further analysis of the expert survey determines the appropriate tax measures that need to be taken in order to

overcome possible consequences for the tax sphere in connection with the development of Industry 4.0 (Table II).

#### IV. DISCUSSION

As our analysis showed, the main technological changes caused by the development of Industry 4.0 are the digitalization of the economy, the use of blockchain technologies, robotics, automation, and M2M technologies (machine-to-machine, data transfer directly between devices), which has a direct impact on changes in the tax sphere.

The digitalization of the economy will manifest itself in various aspects of business sectors, which will affect the development and adaptation of the tax system both in the context of international cooperation and in the context of the development of taxation systems at the national level. Let us consider the typical examples faced in practice by taxation specialists and researchers. We presented all the examples taking into account the rank of the consequences for the economy and taxation received as a result of our study.

##### A. Digital Goods and Services and their Tax Administration

Under the research results [22], industrial enterprises will annually reduce costs by 3.6% and increase revenues by 2.9% over the next five years due to the digitization of products and services, and the development of new digital services. From a tax standpoint, this is a positive trend since income growth also increases tax revenues. However, scholars highlight [13] that digital goods reduce the tax base in several ways: firstly, the cost of digitized goods is lower (for example, books and audio albums); secondly, digital goods and services can be paid for not in cash but in the form of barter (subscription to advertising, newsletter, and other forms of generally B2B communication services); thirdly, digital goods are sold via the Internet (the buyer might be from one country, the seller from another), therefore there is stateless income that is less subject to the current tax laws.

Thus, some countries do not rely on growing revenues from the sale of electronic goods and services of domestic manufacturers and review their taxation systems in order to adapt them to the changes caused by total digitalization. This grants foreign IT companies more access to their markets [26]. However, as the results of our expert survey show, these trends may lead to a decrease in the level of protection of personal and corporate data, so the access of foreign IT companies will be severely limited by the national legislation of countries [27], thereby hindering the development of digital goods and services.

### B. The Use of Tax Instruments as a Way to Protect Personal Data

An additional way to protect information can be the use of tax instruments [28]. At the World Economic Forum of 2011, personal data was recognized as a new asset, whose possession and use can generate income [9]. In the EU, it is allowed to tax enterprises that collect, integrate, and use such data in their activities but this right has not been implemented yet [13]. In France, an attempt was made to introduce a tax on the collection of personal data for Big Data owners (the taxation of Google, Amazon, and Facebook was considered a pilot project) but the corresponding law was not adopted. The reasons include the lack of statistics and schemes for calculating the company's profit from owning such data: on the one hand, their collection and use generate income; on the other hand, it is rather difficult to calculate its share in total income [29].

Researchers [1] believe that work has begun to protect national tax systems and minimize the risks of tax non-payment by digital companies and platforms. One of the proposals is the introduction of specific taxes (a tax on purchases of goods and services via the Internet, or a turnover tax on commercial activities on the web) to prevent the liability of digital businesses.

Currently, additional provisions on the taxation of foreign supplies of digital services and goods are being introduced into the tax codes of some countries. Since 2017, all digital goods and services provided by foreign companies are subject to an indirect tax (Goods and Services Tax) of 10% in Australia; 15% in New Zealand, 8% in Japan; a 5% VAT rate on online purchases in Taiwan [2].

### C. Transfer Pricing of Digital Services

Another feature of the development of Industry 4.0 is transnationalization, therefore an important aspect for tax purposes is the transfer pricing of digital services [19, 30]. According to the experts [17], this can be either quite difficult (if the company's smart connection is installed between a data center located in one jurisdiction and factory floors located in another jurisdiction) or relatively simple (when the intellectual property developed in one jurisdiction (country) is licensed in another jurisdiction).

In the first case, the company's departments where the production facilities are located should, according to transfer pricing rules, pay for the asset (smart connection) the fair market price that a third party could pay for them. However, it is almost impossible to determine the price due to the uniqueness of the asset. Thus, the current transfer pricing models are not always useful. In the second case, when licensing, legal rights to intellectual property usually remain in the country where it was developed, and economic rights are transferred to a foreign jurisdiction. There are no changes in the location of investments: investors use transfer pricing and record the costs of a unit in the country where the property is developed, and the profit is received by a unit in another country that has economic rights [11]. Although it is quite easy to set the price, tax evasion is still possible since, on the one hand, the expenses in the country of the developer (the parent company) will reduce the tax base; on the other hand,

the affiliated party might incur minimum expenses and receive excess profits due to lower tax rates. Therefore, the transfer pricing of companies using new technologies remains open and requires additional research.

### D. Blockchain Technologies' Effect on the Organization of the Taxation System

According to the experts [21], the use of **blockchain technology** will provide internal revenue service field offices with free access to the operations of enterprises as it allows them to simultaneously and automatically calculate tax liabilities, withdraw funds from bank accounts to pay taxes, and eliminate the gap between reporting and paying taxes. Thus, the functions of tax authorities can be significantly reduced, as well as the number of the administrative staff of local tax offices. At the same time, the use of blockchain in order to obtain taxation data will reduce the likelihood of tax disputes and audits.

The first changes might be the elimination of tax returns and the transition to digital tax accounts, which allows one to view and update tax information, receive timely news, and pay tax liabilities. Consequently, society will have a single, centralized digital tax system or platform that will work in real time. Many countries have already been taking steps in this direction. They aim to create a modern and efficient internal revenue service field office, easy to use and with simplified administration in the form of digital taxation.

### E. The Impact of Robotization Processes, Automation, and M2M on Changing the Taxation Organization and System

The introduction of robots into industrial production is quite expensive [31]. To renew assets and maintain the competitiveness of industrial enterprises in the world, a tax credit (R&D) is used [32], which allows for a reduction of the tax base by the amount of an enterprise's costs for the development and implementation of innovations [19].

Besides the advantages of robotics and fully automated production (increase in productivity [33] or wages [34]), there are also disadvantages, in particular, a reduction in the number of jobs, a decrease in demand for low- and medium-skilled workers, an increase in the income gap, i.e. the risks of growing unemployment [12].

According to the study results [15], one of the solutions to growing income disparities can be the redistribution of income with the help of fiscal tools. It is worth mentioning that there are several ways in which such tools can be used for equalization purposes. For example, the introduction of progressive taxes: in the short term, a larger redistribution can be achieved by combining an increase in tax rates for property, the establishment of progressive income taxes, and government programs to support those affected by digitalization and globalization [22].

Another way to address the issue of income inequality with the help of fiscal tools is to introduce a tax on robots [35], i.e. taxing the contribution of robotics and artificial intelligence to the economic results of enterprises. According to scholars [15], such a tax can slow down (at least temporarily) robotics advancement and provide the income



necessary to finance the adaptation of people through retraining programs for dismisses. However, there is an opinion [4] that a serious disadvantage of taxing robots will be the erosion of the tax base and the very concept of robot for tax purposes because this category can include any semi-automatic mechanism. The scholars [4] emphasize that there is a high risk of evading such a tax since elements of robotics will be embedded into mechanisms that are not robots. A possible solution is to create a state trust and introduce universal basic dividends financed by income from the total capital. Thus, an increase in automation and robotics will cause an increase in the income of enterprises that implement them. The automatic distribution of profits in the form of universal basic dividends will be carried out through a state trust that owns a share of such enterprises, which will solve a complex social problem [4].

At the same time, it is impossible to digitalize, introduce robotics and use digital platforms without highly qualified specialists, retraining production personnel, and improving the digital skills of enterprise management [36]. In this regard, we believe it is necessary to introduce loans for students who obtain higher education in the field of science, technology, engineering, and mathematics and are employed in their specialty [16], or an earned income tax credit when taxing the income of teachers who train highly qualified personnel. In Russia, a preferential mortgage program for IT specialists has been developed and is being implemented [37].

## V. CONCLUSION

The study results have confirmed the hypothesis that economic changes as a result of the development of Industry 4.0 technologies affect (and will affect) the tax policy and tax system, which requires the transformation of the modern tax system and the use of both traditional and innovative tax tools.

Transformation of the tax system change the traditional instruments of taxation: the adoption of a progressive tax (for individuals and legal entities); the expansion of taxation objects that emerge due to the digitalization of the economy (electronic goods and services, personal data, Big Data); the introduction of R&D and/or investment tax credit, tax credit for student loans, and preferential mortgage mechanisms for young professionals.

Under the influence of Industry 4.0, innovative tax instruments are being created and introduced into the practice of taxation. In our opinion, the most promising are the introduction of a tax on robots as a tax on the contribution of robotics and artificial intelligence in the economic results of enterprises; taxes on the digital economy (on payments made for the purchase of goods and services via the Internet, or from the turnover from commercial activities on the web).

The limitations of the study include the limited sampling of experts and the geographical representation of experts, as well as the authors' deliberate limitation of focus on technological changes of a certain group: digitalization, blockchain technologies, robotics, automation, and the use of M2M technologies.

This study lacks expert sampling and geographical diversity. In this regard, further research should dwell on the

transformation of taxation as a result of the development of such digital technologies as 3D printing, virtual and augmented reality technologies, digital twin technology, artificial intelligence, Big Data, etc.

## REFERENCES

- [1] E. G. Galizina, A. B. Feoktistova, S. A. Makushkin, I. E. Korotava, E. Y. Kartseva, and N. Udaltsova, "Customer-oriented aggregators of massive open online courses: Ppportunities and prospects", *Webology*, vol. 18, Special Issue, pp. 420-435, 2021.
- [2] R. M. Magomedov, "Digital technologies for competitive analysis and evaluation of competitive capacity of a business entity", *International Journal of Innovative Technology and Exploring Engineering*, vol. 9, no. 1, pp. 1184-1189, 2019.
- [3] I. A. Kiseleva, A. M. Tramova, T. K. Sozaeva, and M. M. Mustaev, "Decision-making modeling in the context of risk and uncertainty caused by social and political processes", *Relacoes Internacionais no Mundo Atual*, vol. 2, no. 34, pp. 44-59, 2022.
- [4] R. V. Batashev, T. E. Zulfugarzade, and A. E. Gorokhova, "Approaches to determining the content and structure of tax administration", *International Journal of Ecosystems and Ecology Science*, vol. 12, no. 3, pp. 385-390, 2022.
- [5] L. P. Grundel, I. A. Zhuravleva, O. V. Mandroshchenko, A. V. Kniazeva, and Y. Y. Kosenkova, "Applications of blockchain in taxation: New administrative ppportunities", *Webology*, vol. 18, Special Issue, pp. 442-443, 2021.
- [6] V. A. Slepov, O. A. Grishina, M. E. Kosov, M. E. Khoranyan, and S. A. Balandin, "Two-parameter model of optimization of the progressive taxation system and its applicability", *Nexo Revista Científica*, vol. 35, no. 1, pp. 412-424, 2022.
- [7] B. Chen, J. Wan, L. Shu, P. Li, M. Mukherjee, and B. Yin, "Smart factory of Industry 4.0: Key technologies, application case, and challenges", *IEEE Access*, vol. 6, pp. 6505-6519, 2018.
- [8] B. Cseh, and J. Varga, "Taxation and humans in the age of the fourth industrial revolution – Financial and ethical", *Acta Universitatis Sapientiae European and Regional Studies*, vol. 17, no. 1, pp. 103-117, 2020. <https://doi.org/10.2478/auseur-2020-0005>
- [9] S. Panasenko, M. Seifullaeva, I. Ramazanov, E. Mayorova, A. Nikishin, and A. M. Vovk, "Impact of the pandemic on the development and regulation of electronic commerce in Russia", *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 5, pp. 652-658, 2022.
- [10] M. Bacache-Beauvallet, and F. Bloch, "Special issue on taxation in the digital economy", *Journal of Public Economic Theory*, vol. 20, no. 1, pp. 5-8, 2017.
- [11] R. Abbott, and B. Bogenschneider, "Should robots pay taxes? Tax policy in the age of automation", *Harvard Law and Policy Review*, vol. 12, pp. 145-175, 2018.
- [12] U. A. Pozdnyakova, A. V. Bogoviz, S. V. Lobova, J. V. Ragulina, and E. V. Popova, "The mechanism of tax stimulation of Industry 4.0 in modern Russia", in *Optimization of the taxation system: Preconditions, tendencies and perspectives*, I. V. Gashenko, Y. S. Zima, and A. V. Davidya, Eds. Cham: Springer Nature, 2019, pp. 189-197.
- [13] C. Frey, and M. Osborne, "The future of employment: How susceptible are jobs to computerization?" *Technological Forecasting and Social Change*, vol. 114, no. C, pp. 254-280, 2017.
- [14] L. Floridi, "Robots, jobs, taxes and responsibilities", *Philosophy & Technology*, vol. 30, no. 1, pp. 1-4, 2017.
- [15] G. Pritchard, D. Hatherell, L. Young, and A. Stocker, *When tax meets technology. Tax implications of Industry 4.0*. Deloitte University Press, 2017, 20 p.
- [16] V. Chand, S. Kostić, and A. Reis, "Taxing artificial intelligence and robots: Critical assessment of potential policy solutions and recommendation for alternative approaches – Sovereign measure: Education taxes/Global measure: Global education tax or planetary tax world", *Tax Journal*, November 2020, pp. 711-761.

- [17] M. Olbert, and C. Spengel, "International taxation in the digital economy: Challenge accepted?" *World Tax Journal*, vol. 1, pp. 3-46, 2017.
- [18] B. McCredie, K. Sadiq, and E. Chapple, "Navigating the fourth industrial revolution: Taxing automation for fiscal sustainability", *Australian Journal of Management*, vol. 44, no. 4, pp. 648-664, 2019.
- [19] V. P. Vishnevsky, and V. D. Chekina, "Robot vs. tax inspector or how the fourth industrial revolution will change the tax system: A review of problems and solutions", *Journal of Tax Reform*, vol. 4, no. 1, pp. 6-26, 2018.
- [20] S. Gupta, M. Keen, A. Shah, and G. Verdier, *Reshaping public finance. Digital revolutions in public finance*. Washington, DC. International Monetary Fund, 2017, pp. 1-21.
- [21] M. A. K. Bahrin, M. F. Othman, N. H. Nor Azli, and M. F. Talib, "Industry 4.0: A review on industrial automation and robotic", *Jurnal Teknologi (Sciences & Engineering)*, vol. 78, pp. 6-13, 2016.
- [22] X. Oberson, "Taxing robots? From the emergence of an electronic ability to pay to a tax on robots or the use of robots", *World Tax Journal*, May 2017, pp. 247-261.
- [23] I. B. Mahalil, A. B. M. Yusof, N. B. Ibrahim, E. M. B. M. Mahidin, N. H. Hwa, "Users' acceptance and sense of presence towards VR application with stimulus effectors on a stationary bicycle for physical training", *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 6, pp. 56-64, 2022.
- [24] S. Dokholyan, E. O. Ermolaeva, A. S. Verkhovod, E. V. Dupliy, A. E. Gorokhova, and V. A. Ivanov, "Influence of management automation on managerial decision-making in the agro-industrial complex", *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 6, pp. 597-603, 2022.
- [25] J. A. Yañez-Figueroa, M. S. Ramírez-Montoya, and F. J. García-Peñalvo, "Measurement of the social construction of knowledge: Validation and reliability of the K-Social-C instrument", *Social Network Analysis and Mining*, vol. 12, no. 1, pp. 50, 2022. <https://doi.org/10.1007/s13278-022-00868-x>
- [26] V. Y. Zhilenko, E. F. Amirova, D. E. Lomakin, N. N. Smoktal, and F. Y. Khamkheeva, "The impact of COVID-19 pandemic on the global economy and environment", *Journal of Environmental Management and Tourism*, vol. 12, no. 5, pp. 1236-1241, 2021.
- [27] V. N. Popov, V. N. Vasilenko, V. A. Khvostov, V. V. Denisenko, A. V. Skrypnikov, A. V. Ivanov, et al. "Security threats to personal data in the implementation of distance educational services using mobile technologies", *Journal of Theoretical and Applied Information Technology*, vol. 99, no. 15, pp. 3935-3946, 2021.
- [28] A. P. Adamenko, A. A. Ananyeva, L. V. Zharapina, I. Y. Tselovalnikova, and J. A. Semenova, "Economic and legal aspects of consumer right protection in tourism", *Journal of Environmental Management and Tourism*, vol. 11, no. 8, pp. 1967-1972, 2020.
- [29] S. Kumar, "Impact of artificial intelligence and biometric authentication on business", *AIP Conference Proceedings*, vol. 23579, 040013, 2022.
- [30] M. A. Kozhevnikova, S. N. Kurbakova, Y. V. Artemyeva, N. V. Palanchuk, and M. M. Umarov, "Development of international tourism in the context of integration processes", *Journal of Environmental Management and Tourism*, vol. 11, no. 4, pp. 1013-1018, 2020.
- [31] A. G. Gurinovich, and M. A. Lapina, "Legal regulation of artificial intelligence, robots, and robotic objects in the field of social relations", *Relacoes Internacionais no Mundo Atual*, vol. 1, no. 34, pp. 55-78, 2022.
- [32] G. V. Kalabukhova, O. A. Morozova, L. S. Onokoy, E. Y. Chicherova, and I. G. Shadskaja, "Digitalization as a factor of increasing investment activity in the tourism industry", *Journal of Environmental Management and Tourism*, vol. 11, no. 4, pp. 883-889, 2020.
- [33] E. V. Potekhina, I. V. Gulina, O. V. Dmitrieva, V. B. Frolova, and J. A. Semenova, "Increasing labor efficiency in the area of digital entrepreneurship", *Revista Inclusiones*, vol. 7, no. Especial, pp. 137-145, 2020.
- [34] D. Janteliyev, T. Julamanov, B. Rsybetov, A. Kaldybekov, and Y. Allaberganova, "Increasing the level of management efficiency: Using unmanned aerial vehicles for monitoring pasture lands", *Instrumentation Measure Metrologie*, vol. 21, no. 2, pp. 59-65, 2022.
- [35] A. G. Gurinovich, and N. I. Petrykina, "Specifics of developing the institution of public service: International experience and its application in Russia", *Jurídicas CUC*, vol. 17 no. 1, pp. 253-276, 2021.
- [36] A. Neznamova, G. Kuleshov, and M. Turkin, "International experience in personal data protection", *Jurídicas CUC*, vol. 16, no. 1, pp. 391-406, 2020. <http://dx.doi.org/10.17981/juridcuc.16.1.2020.17>
- [37] G. Trushin, *Russia has launched preferential mortgages for IT professionals. How it works*. 2022. [Online]. Available: <https://realty.rbc.ru/news/624c68939a7947787d21777>

# Attractiveness of the Megaproject Labor Market for Metropolitan Residents in the Context of Digitalization and the Long-Lasting COVID-19 Pandemic

Mikhail Vinichenko<sup>1</sup>, Sergey Barkov<sup>2</sup>, Aleksander Oseev<sup>3</sup>, Sergey Makushkin<sup>4</sup>, Larisa Amozova<sup>5</sup>

Russian State Social University, Moscow, Russia<sup>1,4</sup>  
Lomonosov Moscow State University, Moscow, Russia<sup>2,3,5</sup>

**Abstract**—The article aims to determine the nature of changes in the attractiveness of the labor market of megaprojects from the perspective of megapolis residents under the conditions of digitalization and the long-lasting pandemic of COVID-19. The paper develops a scientific-methodological and categorical-conceptual apparatus with the support of empirical methods with distance methods. The study shows that the attractiveness of the labor market of megaprojects has undergone certain changes for megapolis residents under the current conditions. The factors of the attractiveness of the labor market of megaprojects are of a stable nature in the minds of megapolis residents. The main advantage of the work is the identification of trends in the changes of the megaproject labor market and the relationships they have. The study reveals both general and private trends. The obtained results can be used for further study of the megaproject labor market and the improvement of the social policy of the state and megalopolises in the conditions of digitalization and the prolonged pandemic.

**Keywords**—Megaproject labor market; metropolitan residents; digitalization; COVID-19 pandemic; attractiveness factors

## I. INTRODUCTION

The fourth industrial revolution (4IR) and the digitalization of all spheres of life are conceptually changing the structure of the labor market and the system of socio-economic relations. The more rapid the digitalization, the more contradictions arise and the greater the resistance from certain segments of society. The complexities associated with the COVID-19 pandemic are added to this tangle of contradictions. The world economy is undergoing structural changes with an emphasis on mass digitalization and remote forms of employment. The global recession hits small and medium-sized businesses the hardest, service businesses suffer losses [1]. Transforming the market for goods and services, the introduction of artificial intelligence reduces customer confidence, worsens the labor market [2], and furthers social inequalities. The coronavirus pandemic worsens the plight of workers, increases their dependence on employers and the digital environment.

Megaprojects are implemented under special conditions. Recently, large projects are mainly social and concern water supply, social housing [3], and infrastructure development. Their effectiveness in a digital environment depends on the

management system that is created [4, 5], the adopted decision-making procedures, timing forecasting technologies [6], risks, and funding procedures. With weak management, megaprojects threaten regional ecology [7] and local communities [8] and conflict with the Sustainable Development Goals (SDGS). Amid the COVID-19 pandemic, several projects aim to curb the dangerous virus, while others try to make super profits by degrading the environment and the living conditions of the population. Such approaches adversely affect the labor market of megaprojects.

The introduction of digital services and artificial intelligence into the sphere of human resources (HR) only partially compensates for the effect of negative factors on the labor market. Innovative technologies based on digitalization contribute to the development of social partnerships [9], especially in megapolises. Young people find themselves in a contradictory situation. On the one hand, they are the most adapted to digitalization, constant change, and the application of themselves in new areas of business. They easily enter the labor market with the help of social networks, quickly obtain information about all kinds of changes. Young professionals change their views on employment in the international labor market, on the forms and content of work itself [10], their behavior in the workforce. Under these conditions, it is necessary to develop mechanisms to attract the attention of young people, to improve their motivation [11], and to create favorable working conditions for entering megaprojects [12]. Higher education [13] and the role and professionalism of recruiters are growing in importance. IT professionals are in a better position, as the demand for them is only growing in both megaprojects and megapolises.

At the same time, in the context of the pandemic, it is difficult for young professionals to compete with experienced workers in large companies who have earned the trust of employers and have a strong hold on their jobs. While at the beginning of the pandemic, it was mostly young people who sought remote work, by the second year, more and more age-matched residents of large cities had joined the ranks. Unemployment in various sectors of the economy is growing, which to some extent encourages young people and experienced workers to seek jobs in megaprojects in remote areas.

The acuteness of the problem of providing megaprojects with personnel has not yet led to the development of a precise method for promoting the attractiveness of the megaproject labor market among megapolis residents. This process is further complicated by the prolonged COVID-19 pandemic. The present paper may become an element in the system of measures to improve the effectiveness of megaprojects by means of competent and technologically advanced provision of the workforce for them with a focus on megapolis residents.

## II. MATERIALS AND METHODS

### A. Design and Hypothesis

The present study is part of a research project and the next stage in identifying the nature of the impact of the digitalization of the economy and the use of artificial intelligence on the social environment and labor market in general [14], and on the labor market of megaprojects in particular. The peculiarity of this work is that it takes into consideration the impact on the labor market of such a negative factor as the limitations of the pandemic. The work develops a scientific-methodological and categorical-conceptual apparatus, which allows achieving the goal of the study by solving scientific problems. The study puts forward hypotheses and develops a set of approaches and methods in the combination developed specifically for this study. The research team is formed with a specific allocation of tasks and functions, the order of the research is determined. The priority of the research team is the observance of scientific ethics in conducting the research.

The goal of the study is to determine the nature of changes in the attractiveness of the megaproject labor market from the perspective of megapolis residents under the conditions of digitalization and the long-lasting COVID-19 pandemic. To achieve the goal, the following research objectives are set for the study:

1) To determine the degree of readiness of megapolis residents to enter the megaproject labor market in the context of the digitalization of society and the prolonged pandemic of COVID-19.

2) To reveal the essence of changes in megapolis residents' assessments of megaproject attractiveness factors.

The hypotheses proposed by the team of authors are as follows:

H1. The attractiveness of the megaproject labor market has undergone significant changes for megapolis residents in the context of digitalization and the long-lasting pandemic of COVID-19.

H2. Megapolis residents' assessment of the megaproject labor market is contingent on their sex, age, and sphere of work and is predictable.

The comprehensive study consists of two stages: the first stage lasting from January 10, 2020, to June 10, 2021, with a sample of n=719 people (2021.1) and the second stage from July 1 to December 30, 2021, with n=1098 (2021.2) people with a general population of n=12'500'000. The sampling error is 3.75% with a 95% confidence level for the first stage

of the study and 3.5% with a 95% confidence level for the second stage. The main quota characteristics in the study when selecting respondents are gender, age, level of education, and work experience.

The decision to organize two stages of the study (two studies) is due to the possibility of differences in the views of megapolis residents at the start of the pandemic after a certain cycle of it and in the context of the prolonged pandemic, which has led to moral and psychological exhaustion, deterioration of health, pressure on the part of the government and the employer management, changes in the labor market, and the transformation of the structure of the market of goods and services.

Analysis of socio-demographic characteristics of the respondents (Table I) shows that in both studies, women are more active than men in almost the same ratio. It is natural both in view of the general ratio of the male and female population of the country and the higher activity of women in sociological surveys.

The age ratio of the respondents differs between the two stages to some degree. The proportion of metropolitan residents who participated in the sociological survey under the age of 25 has increased (from 67% to 76%). At the same time, the number of respondents under 18 years old has significantly increased (from 2% to 27%). Analysis of other age groups shows that the older generation, even among young people, is tired of numerous sociological surveys conducted on various occasions (health care, banking, political priorities...) by various state and commercial structures.

TABLE I. SOCIO-DEMOGRAPHIC CHARACTERISTICS OF THE RESPONDENTS (IN %)

Characteristics of respondents		Share	
		2021.1	2021.2
Sex	male	31%	30%
	female	69%	70%
Age	14 - 18 years old	2%	27%
	19 - 25 years old	65%	49%
	26 - 35 years old	22%	18%
	36 - 55 years old	9%	3%
	56 - 65 years old	1%	2%
	over 65 years old	1%	1%
Level of education	higher	39%	15%
	incomplete higher	38%	35%
	secondary special	15%	18%
	secondary	7%	31%
	elementary	1%	1%
Work experience	1 year	26%	68%
	1-3 years	35%	20%
	4-5 years	18%	7%
	6-10 years	11%	2%
	over 10 years	10%	3%
Nature of work activities	executive	13%	4%
	specialist	48%	59%
	government official	11%	14%
	blue-collar worker	15%	12%
	self-employed	9%	8%
	unemployed	4%	3%

The age shift also affects the level of education. The share of respondents with higher education is significantly lower

(from 39% to 15%), while the group with secondary education has grown (from 7% to 31%). The latter are interested in discussing their future with the opportunity to participate in megaprojects as a mechanism of upward social mobility. The labor market itself, in turn, is in dire need of a great number of specialists with specialized secondary education. The provision of the market with highly qualified personnel is also a problematic issue.

Work experience in proportion to age has shifted toward one year. A significant reduction is observed in the number of managers (from 13% to 4%) and respondents with over four years of experience (from 21% to 5%). This suggests that experienced employees and managers have become less active in discussing their future in terms of participation in megaprojects. They are more comfortable working in their positions in the metropolitan city of Moscow.

### B. Data Collection and Sample

The study is organized and conducted collaboratively by the Humanities Department of the Russian State Social University (RSSU) and the Department of Economic Sociology and Management of the Sociology Department at the Lomonosov Moscow State University (MSU).

The research methodology is based on the sequential use of a group of methods, beginning with empirical ones. The data obtained in the course of empirical research serves as a basis for statistical processing, systematization, comparative analysis, and discussion.

First, a questionnaire was prepared for a sociological survey, which underwent expert evaluation. The indicators are assessed using the Likert scale. The scaling went through five levels, from complete disagreement to complete agreement with the definition. Due to the pandemic restrictions, the sociological survey was conducted remotely using special programs (Google Form), cloud conference platform Zoom, and VoIP service Skype.

The cause-and-effect relations on the problems of selecting priority attractiveness factors and their antipodes are discovered in the course of in-depth interviews. The respondents for the in-depth interview are selected randomly. In the first study, the in-depth interview involves 18 respondents, in the second study – 21 respondents. The content and structure of the in-depth interview are developed taking into account the results of the conducted surveys.

Systemic conclusions to determine the nature of changes in the attractiveness of the megaproject labor market for megapolis residents in the conditions of digitalization and the prolonged COVID-19 pandemic are developed based on a focus group. The focus group consists of Russian and foreign experts in the field. In the first study, the focus group includes eight experts, and in the second study – nine experts.

In light of the changes in the social environment due to the digitalization of society in the context of the long-lasting COVID-19 pandemic, the categorical-conceptual apparatus on the issue is developed (finalized).

The term “megaproject” is composite. It refers to a set of projects that are aimed at achieving a specific goal,

interconnected by tasks, place, and time, and provided with the necessary resources. In this study, under the megaproject labor market, we understand the nature and totality of the supply of jobs by the composite employer uniting the set of organizations that offer jobs as part of a large project (megaproject) at all stages of its realization, as well as the totality of demand for those jobs by potential participants in the project. Due to the highest concentration of labor resources of different specializations and qualifications in megapolises, the paper will consider the possibility of their participation in the planned and ongoing Russian and foreign megaprojects. Digitalization of society is the process of introducing information and communication, digital technologies in all socio-economic structures and spheres of life. By artificial intelligence (AI), we understand intelligent computer programs, systems tasked with creating intelligent reasoning and actions. Sometimes, AI is designed to look like robots, including human appearance [15].

The attractiveness of megaprojects is determined by attractiveness factors, which include the most significant factors that ensure the desire (aspiration) of megapolis residents to participate in the megaproject. On the eve of the sociological survey, the expert group ranked the attractiveness factors to include them in the questionnaire. For this purpose, a matrix of pairwise comparisons is used. The sample consists of the 10 most significant factors, including such factors as gaining work experience in a large project, interest in communication with different people, opportunity to make a career, as well as receiving a good salary.

The draft of the survey is tested in a pilot study on the sample of the citizens of Moscow living in the Ostankinskiy District of the North-Eastern Administrative Okrug of the city. This study used the “snowball”, the respondents passed on the invitation to participate in the sociological survey through their social communication channels.

### III. RESULTS

The study demonstrates that the ratio of megapolis residents willing to enter the megaproject labor market and participate in megaprojects remains the same. The greatest part of the respondents is to a certain extent ready to leave the megapolis (or stay in it) to participate in megaprojects (Fig. 1).

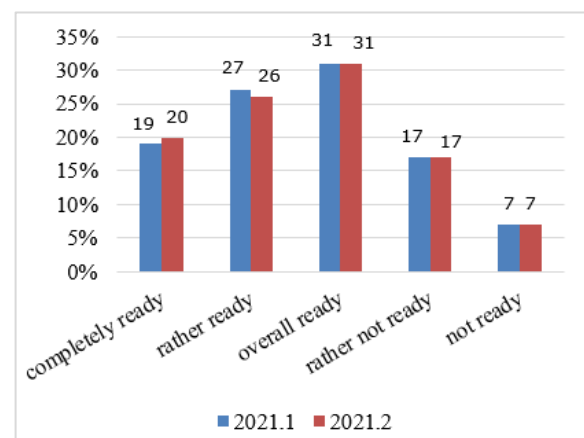


Fig. 1. Response Options to the Question, “Are you Ready to Offer Yourself in the Megaprojects Labor Market?” Source: Own Research, 2021.

It should be noted that the duration of the pandemic and the desire of the leadership to digitalize the private part of the people does not produce a significant effect on the position of megapolis residents. The share of the respondents completely ready to change the place of work (study) and take part in the megaprojects remains within 19-20%, and categorically refuse such opportunities the same 7% of the respondents. The rest of

the sample fluctuates between being partially ready or partially not ready to participate in megaprojects (57-58%).

The factors of attractiveness proposed for evaluation also remained approximately within the same limits of expert preferences. The changed conditions of the digitalization of society and the long pandemic have only partially changed the desires and aspirations of megapolis residents (Fig. 2).

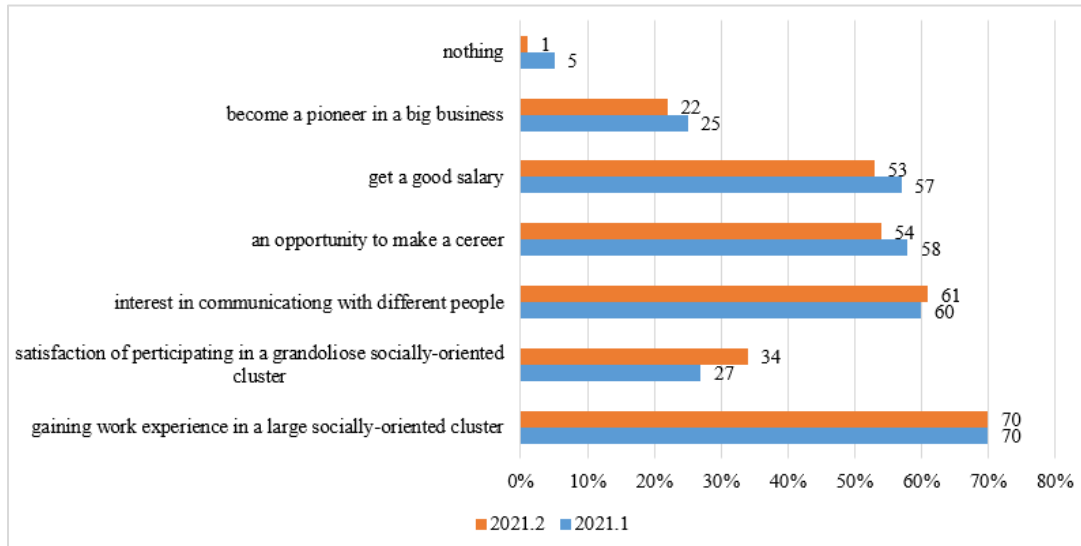


Fig. 2. Response Options to the Question, "What Attracts you most to a Megaproject?" Source: Own Research, 2021.

Among the leading **attractiveness factors** are the opportunity to gain experience working in a major project (70% of respondents), interest in interacting with different people (60-61%), the opportunity to make a career (54-58%), and getting a good salary (53-57%). Some changes are observed in the evaluation of the importance of other attractiveness factors. The importance of satisfaction with participation in a large megaproject is slightly decreased (from 34% to 27%). Total denial has increased with respect to the importance of megaprojects (from 1% to 5%), as well as of becoming a pioneer in a big business (from 22% to 25%).

At the bottom of the table of attractiveness factors, with less than 1% of the vote are the same factors, but with some changes in the values (Table II).

TABLE II. FACTORS WITH A LOW LEVEL OF ATTRACTIVENESS IN THE LABOR MARKET MEGAPROJECTS (IN %)

n/n	Factors of the attractiveness of a megaproject	%	
		2021.1	2021.2
1	Implementation of professional skills	0.1	0.1
2	If the company is international – the opportunity to communicate in a foreign language	0.1	0.2
3	Access to information	0.1	0.1
4	Any project is better than deployment in the army	0.1	0.2

At the same time, slightly higher is the importance of the opportunity to communicate and gain experience in a foreign language in an international company (from 0.1 to 0.2%) and to take part in a megaproject instead of serving in the army (from 0.1 to 0.2%).

In general, the study reveals the nature of the views of the residents of the Moscow metropolitan area on the megaproject labor market in the context of the digitalization of society and the prolonged pandemic.

#### IV. DISCUSSION

Comparative analysis of the sociological surveys reveals that the desire to enter the labor market of megaprojects generally remains unchanged among megapolis residents. However, taking into account changes in the age composition of respondents, their work experience, and professional skills, we can conclude that young people under 25 years old, mostly with secondary, specialized secondary, and incomplete higher education, are more favorable to megaprojects. They endured the pandemic digitalization with the least losses, being in educational institutions for the most part of it. The share of opponents of participation in megaprojects remains consistently small (7%).

The study also demonstrates a great number of those in doubt (over half of the respondents). This is explained by the instability of the situation: the economic crisis, the ever-changing restrictions, contradictory actions on the part of leadership to combat the COVID-19 pandemic, problems for small and medium-sized businesses, and increasing demands for skills and digital competencies on the part of employers. This group of respondents above all needs guarantees of the stability of employment in a big company in a decent position, as well as favorable working conditions. This is consistent with several studies [16].



The focus group experts are also sure of the hypothetical readiness of megapolis residents to enter the megaproject labor market under the condition of a high salary. This, however, applies mostly to young people with modest professional skills. Executives and highly qualified specialists will offer their services in this labor market reluctantly and involuntarily. Middle-aged and older respondents are least inclined to change the atmosphere of the Russian capital for dubious prospects in megaprojects. This is especially true for IT, logistics, and marketing specialists. These specialists were able to adapt to the constantly changing conditions and requirements of the megapolis administration and company management in the rich metropolitan labor market. They have developed several measures and implemented solutions to bypass (overcome) the next pandemic restrictions on movement, forced-voluntary vaccination, and remote labor and social activities. At the same time, they have the opportunity to partially or fully participate in megaprojects remotely, without leaving the megalopolis. The digitalization of the socio-economic sphere allows this to happen. This confirms the general **trend** of fragmented hypothetical desire and remote participation of megapolis residents in megaprojects.

Here it needs to be pointed out that a part of megapolis residents of various ages, being driven to despair by restrictions and oppression, openly and sometimes aggressively oppose such restrictions.

The in-depth interview reveals cause-effect relationships in determining the priority of attractiveness factors. The desire of megapolis residents to enter a new labor market to gain experience in a major project appears to be stable and strong. This seems logical for young people since, in the course of the pandemic, older workers have developed mechanisms for survival and retention of their positions. In megaprojects young professionals see an opportunity to rapidly develop their career, realize their creative potential, get a good salary, raise their social status, and get an opportunity to join a promising ambitious team. These findings correlate with several studies.

Work in a megaproject usually involves interaction with a large number of workers of different professions and areas of work. This arouses steady interest among megapolis residents. At the in-depth interview, they express their hopes to work with a good manager, in a favorable social and psychological climate with a friendly attitude to all employees regardless of their age, gender, social status, socio-ethnic features, and religious affiliation. The importance of these aspects in attracting and retaining employees in companies has been emphasized by various researchers [17, 18].

At the same time, in-depth interviews express concerns about entering the process of mass recruitment, the fear of "getting lost" in large teams, the difficulty of adaptation, especially on the part of talented young people.

The respondents' desire for good salaries is based on considerable and usually sustainable financing of large projects. This also implies high salaries in all positions of the megaproject. It should be noted that the concept of good wages among the megapolis residents differs significantly

from the views of the regional population. This acts as a deterrent to attracting megapolis labor resources to megaprojects. There is also a certain differentiation in the level of wages by age. For the majority of young people, the requests are lower than those of older age groups. These aspects are reflected in the works of researchers exploring the problem of personnel **motivation**.

The ambitiousness of work in the projects remains generally at the same level. For the most part, young people express their desire to be pioneers in the implementation of large-scale digital projects that would significantly advance the digitalization of the Russian economy. To a certain extent, the respondents point to the pandemic restrictions being lighter in the remote regions that could potentially become the site of the megaproject. However, these statements do not carry much meaning. Youthful enthusiasm, the desire to create the groundwork for a great career remain among the priorities.

The focus group experts conclude that in the second stage of the study, the main attractiveness factors have the greatest effect on the age group under 25 years old. This points to the emergence of a **private trend** of the decrease in age in assessing the effectiveness of the impact of the factors of the attractiveness of the megaproject labor market from the perspective of megapolis residents under the conditions of digitalization and the prolonged COVID-19 pandemic. At the same time, the results support the discovered **dependence** of the labor activity of megapolis residents in megaprojects on their age: the younger the age, the higher the communicability, labor activity, and readiness to relocate and change the place and conditions of work. This dependence was taken into account by the leadership of the Soviet state in the design and successful implementation of megaprojects.

The opportunities to apply professional skills in the megaproject and have access to its information remain in the zone of low demand. This is especially characteristic of young people who have only basic professional skills and limited work experience that had little to do with digital services and was rather a part-time job in the service sector in low-skilled positions, offline distribution of advertising. Focus group experts note young people's desires and simultaneous fears with respect to mastering new digital technologies and being socially secure. To some extent, this falls in line with the works of other researchers [10, 19, 20]. Somewhat higher is the share of proponents of communicating in a foreign language and military evaders. This comes as a result of the opportunity to gain greater career prospects by knowing a foreign language, as well as of a tougher approach to military personnel in the fight against the pandemic in the army.

Overall, it can be argued that there has emerged a common downward **trend** in the age of metropolitan residents willing to enter the megaproject labor market in the face of digitalization and the long pandemic of COVID-19.

## V. CONCLUSION

The study establishes that the attractiveness of the labor market of megaprojects has undergone certain changes for megapolis residents under the conditions of digitalization and the long-lasting pandemic of COVID-19. This partially

confirms the first hypothesis. The essence of these changes lies in the plane of age and the sphere of work. The work confirms the dependence of megapolis residents' employment in megaprojects on their age revealed in the first stage of the study: the younger the person, the higher their communicativeness and labor activity, readiness to relocate and change the place and conditions of work. To a certain extent, this confirms the second hypothesis that megapolis residents' assessment of the megaproject labor market depends on gender, age, and sphere of work and has a predictable nature.

Attractiveness factors have not undergone major changes in terms of significance, which points to the validity of the conducted expert assessments and the stability of opinions of megapolis residents. Those megapolis residents having real opportunities to enter the rich labor market of the capital are steadily attracted to megaprojects by the opportunity to gain experience in a major project, the interest of communicating with different people, the opportunity to make a career, and a high salary.

The study reveals (proves) general and private trends. Among the discovered **general trends** are the reducing age of megapolis residents wishing to enter the megaproject labor market in the context of digitalization and the prolonged COVID-19 pandemic, as well as the trend of fragmented hypothetical desire and remote participation of megapolis residents in megaprojects. Of private nature is the **trend** of the reduction of age in assessing the effectiveness of the impact of factors in the attractiveness of the megaproject labor market.

The results of the comparative analysis may be of interest for further research on the peculiarities of socio-cultural and socio-economic life in the context of the long-lasting COVID-19 pandemic with the concentration of digitalization on individual social aspects (QR codes). Furthermore, the results of the analysis may be used in covering the staffing needs of emerging and existing megaprojects, to be attentive to people, and to improve the social policies of the state and megapolises.

#### REFERENCES

- [1] K. F. Zimmermann, G. Karabulut, M. Huseyin Bilgin, M., and A. Cansin Doker, "Inter-country distancing, globalization and the coronavirus pandemic", *The World Economy*, vol. 43, no. 6, pp. 1484-1498, 2020. <https://doi.org/10.1111/twec.12969>
- [2] Z. Tong, H. Chen, X. Deng, K. Li, and K. Li, "A scheduling scheme in the cloud computing environment using deep Q-learning", *Information Sciences*, vol. 512, pp. 1170-1191, 2020. <https://doi.org/10.1016/j.ins.2019.10.035>
- [3] A. Garay, A. Ruiz, and J. Guevara, "Dynamic evaluation of thermal comfort scenarios in a Colombian large-scale social housing project", *Engineering, Construction and Architectural Management*, vol. 29, no. 5, pp. 1909-1930, 2021. <https://doi.org/10.1108/ECAM-09-2020-0684>
- [4] E. Hetemi, A. van Marrewijk, A. Jerbrant, and M. Bosch-Rekvelde, "The recursive interaction of institutional fields and managerial legitimation in large-scale projects", *International Journal of Project Management*, vol. 39, no. 3, pp. 295-307, 2021. <https://doi.org/10.1016/j.ijproman.2020.11.004>
- [5] J. Larumbe, J. Garcia-Barruetaña, and D. Lopez De Ipiña-Gonzalez De Artaza, "Methodology for the implementation of configuration management on large scale projects", *Dyna*, vol. 96, no. 1, p. 13, 2021. <https://doi.org/10.6036/9806>
- [6] M. Gordon, D. Viganola, A. Dreber, M. Johannesson, and T. Pfeiffer, "Predicting replicability-analysis of survey and prediction market data from large-scale forecasting projects", *PLoS ONE*, vol. 16, no. 4, e0248780, 2021. <https://doi.org/10.1371/journal.pone.0248780>
- [7] M. V. Rybakova, M. V. Vinichenko, Y. S. Ushakova, O. L. Chulanova, S. A. Barkov, M. A. Malyshev, et al., "Ecological problems of Russian cities on the views of young people", *Ekoloji*, vol. 28, no. 107, pp. 5019-5026, 2019.
- [8] Z. Wang, R. Nixon, A. Erwin, and Z. Ma, "Assessing the impacts of large-scale water transfer projects on communities: Lessons learned from a systematic literature review", *Society and Natural Resources*, vol. 34, no. 6, pp. 822-843, 2021. <https://doi.org/10.1080/08941920.2020.1859029>
- [9] N. V. Medvedeva, E. V. Frolova, and O. V. Rogach, "Territorial public self-government and local government: Interaction and prospects for partnership", *Sotsiologicheskie Issledovaniya*, vol. 10, pp. 72-82, 2021. <https://doi.org/10.31857/S013216250015275-5>
- [10] T. S. Demchenko, P. Karácsony, I. Y. Iliina, M. V. Vinichenko, and A. V. Melnichuk, "Self-marketing of graduates of high schools and young specialists in the system of personnel policy of the organization", *Modern Journal of Language Teaching Methods*, vol. 7, no. 9, pp. 58-65, 2017.
- [11] O. L. Chulanova, O. L. Ryngach, M. V. Vinichenko, O. V. Kaurova, M. V. Demchenko, and T. S. Demchenko, "Increase of staff loyalty by improving the motivation (stimulation) system in enterprises oil and gas complex of the Khanty-Mansiy autonomous district - Ugra", *Modern Journal of Language Teaching Methods*, vol. 8, no. 8, pp. 303-314, 2018.
- [12] O. V. Rogach, E. V. Frolova, A. V. Kirillov, V. V. Bondaletov, and M. V. Vinichenko, "Development of favourable learning environment and labor protection in the context of harmonization of social interaction of educational system objects", *Mathematics Education*, vol. 11, no. 7, pp. 2547-2558, 2016.
- [13] E. V. Frolova, and O. V. Rogach, "Digitalization of higher education: Advantages and disadvantages in student assessments", *European Journal of Contemporary Education*, vol. 10, no. 3, pp. 616-625, 2021.
- [14] M. V. Vinichenko, O. L. Chulanova, M. V. Rybakova, S. A. Barkov, and M. A. Malyshev, "The impact of artificial intelligence on behavior of people in the labor market", *Journal of Advanced Research in Dynamical and Control Systems*, vol. 12, no. 4 (Special Issue), pp. 526-532, 2020. <https://doi.org/10.5373/JARDCS/V12SP4/20201518>
- [15] M. V. Vinichenko, M. V. Rybakova, O. L. Chulanova, S. A. Barkov, S. A. Makushkin, and P. Karacsny, "Views on working with information in a semi-digital society: Its possibility to develop as open innovation culture", *Journal of Open Innovation: Technology, Market, and Complexity*, vol. 7, no. 2, pp. 160, 2021. <https://doi.org/10.3390/joitmc7020160>
- [16] M. V. Vinichenko, E. V. Frolova, E. E. Kabanova, M. S. Kozyrev, and T. A. Evstratova, "The youth employment problems", *Journal of Advanced Research in Law and Economics*, vol. 7, no. 2, pp. 378-387, 2016.
- [17] G. Nikiporets-Takigawa, "Youth and youth policy in the UK: Post-brexite view", *Sovremennaya Evropa*, vol. 1, no. 80, pp. 47-58, 2018.
- [18] A. A. Oseev, F. A. Dudueva, P. Karácsony, M. V. Vinichenko, and S. A. Makushkin, "The peculiarity of the ethno-social conflicts in the Russian labor market: Comparative analysis of Russia, Great Britain and Germany", *Espacios*, vol. 39, no. 22, p. 12, 2018.
- [19] E. Nkansah-Dwamena, and A. Bonnie Raschke, "Justice and fairness for Mkangawalo people: The case of the Kilombero Large-scale Land Acquisition (LaSLA) Project in Tanzania", *Ethics, Policy and Environment*, vol. 24, no. 2, pp. 137-163, 2021. <https://doi.org/10.1080/21550085.2020.1848187>
- [20] L. Zhongyuan, A. V. Melnichuk, S. A. Makushkin, M. V. Vinichenko, and M. A. Azhmuratova, "Social protection management: Anti-fraud technologies", *Journal of Advanced Research in Dynamical and Control Systems*, vol. 12, no. 4, pp. 1687-1693, 2020. <https://doi.org/10.5373/JARDCS/V12SP4/2020165>

# Generation and Assessment of Intellectual and Informational Capital as a Foundation for Corporations' Digital Innovations in the "Open Innovation" System

Viktoriya Valeryevna Manuylenko<sup>1</sup>, Galina Alexandrovna Ermakova<sup>2</sup>, Natalia Vladimirovna Gryzunova<sup>3</sup>, Mariya Nikolaevna Koniagina<sup>4</sup>, Alexander Vladimirovich Milenkov<sup>5</sup>, Liubov Alexandrovna Setchenkova<sup>6</sup>, Irina Ivanovna Ochkolda<sup>7</sup>

North Caucasus Federal University, Stavropol, Russia<sup>1,2</sup>

Russian Technological University, Stavropol Branch, Stavropol, Russia<sup>1</sup>

Plekhanov Russian University of Economics, Moscow, Russia<sup>3</sup>

Saint Petersburg State Marine Technical University, Saint-Petersburg, Russia<sup>4</sup>

Moscow Economic Institute, Moscow, Russia<sup>5</sup>

Financial University under the Government of the Russian Federation, Moscow, Russia<sup>6</sup>

International University MITSO, Affiliate in Gomel, Gomel, Belarus<sup>7</sup>

**Abstract**—The research peruses development of scientifically based toolkit to create and assess promising types of intellectual capital transformed into digital innovations for "open innovation" system. It is determined that in theory terms "intellectual, informational and digital capitals" are interrelated categories; efficient merge of informational and digital capitals minimizes information security risks; merge of informational and digital capitals provides a long-term multiplicative synergetic effect demonstrating constant transformation of innovative ideas into digital innovations. The following is suggested: structural and logical scheme method for creation and assessment of informational capital and scenarios for the synergetic development of informational and digital capitals.

**Keywords**—*Informational and digital capitals; informational security risks; synergy; open innovation; transformation*

## I. INTRODUCTION

In order to keep and maintain high financial corporate status, commercial organizations (corporations) should constantly develop promising trends of intellectual capital generation and its application according to the principle "intellectual capital available in corporation means future potential innovations" in the "open innovation" system. Modern digitalization of business processes calls for digital innovations development that first of all requires intellectual, informational and digital capitals synergy in the "open innovation" system. Timely transformation of intellectual, informational and digital capitals into digital innovations is hampered by the lack of inclusive understanding of the "informational and digital capitals" essence, as well as the processes of their generation, assessment and application. Poor methodological and practical background for creation of a toolkit for generating and assessing intellectual, informational and digital capitals - which could provide for development of digital innovations - limits corresponding

transformation. In this regard, substantiation of theoretical and methodological provisions and practical trends for creation and development of a toolkit for generation and assessment of intellectual informational and digital capitals with its subsequent transformation into digital innovations in the "open innovation" system is a key issue for foreign and national science and practice. Therefore, it determined relevance of the current research topic.

Achievement of the research goal suggests the following tasks: investigation of the essential characteristics of concepts - "intellectual informational capital" and "intellectual digital capital" identifying cross links; consider possibility to transform intellectual, informational and digital capitals into digital innovations in the "open innovation" system.

Working hypothesis of the study is based on the need to develop scientifically based toolkit for intellectual, informational and digital capitals generation and their assessment according to the principle, stipulating that intellectual capital available in corporations means future potential innovations" in the "open innovation" system, which is aimed at further digitalization of business processes in the moment in time and for long-term perspective. Theoretical significance of the study lies in deeper and more expanded understanding of intellectual, informational and digital capitals' generation and assessment, their transformation into digital innovations in the "open innovation" system. Some of the theoretical and methodological aspects of the research are proposed as tutorial and methodological materials for some disciplines of educational programs, as well as for retraining programs and advanced training of the corporations' employees working in the assessment, financial and innovation spheres, etc. Practical significance of the research lies in development and application of specific methods, models and practical provisions that serve to shape

methodological and practical basis for educational processes, assessments of intellectual, informational and digital capitals, and transformation of the latter into digital innovations; that will determine future implementation of a new methodological toolkit that ensures development and promotion of corporations' digital innovations in the "open innovation" system.

The role of intellectual capital in stimulating the transition of healthcare networks to digital technologies is reflected very well in the study of F. Schiavone et al. [1]. In particular they studied creation of an information panel to monitor key performance indicators (KPIs) in digital healthcare networks. It is aimed at improving health policy by means of developing an integrated meso-level structure based on the central role of intellectual capital components (structural, relational, human). In the context of pandemic uncertainty, which demonstrated the need for digital information technologies, the study of K.S. Al-Omouh, D. Palacios-Marqués, and K. Ulrich [2] on the interaction of intellectual capital, supply chain flexibility, collaborative knowledge creation and corporate sustainability during unprecedented crises such as Covid-19 is of high interest. The data was analyzed with the software for modeling structural equations with partial least squares (Smart-PLS). S. Kusi-Sarpong et al. [3] noted the connection of intellectual capital, including human, structural capital, and digital technologies capital - blockchain, embedded in supply chain management that ensures sustainable production.

In the study [4], special attention is paid to determination of the investments impact into intellectual capital as they describe the trends of industrial revolution 4.0 and intellectual capital in the era of FinTech.

E. García-Meca and I. Martínez [5] recognize how information influences intellectual capital while investment decisions making, and point out that certain characteristics of the firm affect the use of intellectual capital information.

Intellectual capital plays strategic role in achieving sustainable development goals and defining future, as noted by G. Secundo, V. Ndou, P.D. Vecchio and G. De Pascale [6]. A. Kalkan, Ö.Ç. Bozkurt and M. Arman [7] point to the connection of intellectual capital, innovation, organizational strategy and firm efficiency, revealing the impact of intellectual capital, innovation and organizational strategy on the activities of the companies operating in Antalya, Turkey.

F. Ricci, V. Scafarto, S. Ferri and A. Tron [8] insist that disclosure of information related to digitalization is a form of disclosure of information about intellectual capital that provides investors with potentially valuable data.

Taking into account the above, recognition of the connection between intellectual digital and information capitals becomes clear as it is the basis for digital innovations development in corporations in the "open innovation" system, which must be constantly improved and developed in the conditions of economic processes digitalization.

## II. METHODOLOGY

### A. Identification of Intellectual Capital and Innovations Cross Links

Considering that the best method for measuring intellectual, informational and digital capitals, which are the basis for digital innovations' development in corporations in the "open innovation" system, is the method that reflects their economic content and its generation is carried out in the following sequence of stages:

Stage 1: identification of the connection between intellectual capital and innovations. Its result is identification of the resource for innovative activity in corporations in the future, i.e. their intellectual capital.

Stage 2: ensuring the synergy of intellectual capital assessment and innovations, which is based on the principle stating that "intellectual capital available in the corporation is its future innovations". Its result is the determination of the factors for productive reproduction of intellectual capital that limit innovations including organizational innovations.

Stage 3: alternative direction for interpretation and evaluation of the intellectual information capital in corporations. Its result is provision of the characteristics for intellectual information capital in corporations, which is designed to ensure informational security of intellectual activity.

Stage 4: interpretation of the intellectual digital capital in corporations, its interaction with human and informational capital. Its result is justification of the fact that digital innovations represent interconnection of human, informational and digital capitals in "open innovation" system.

Theoretical and methodological foundation of the research: works of the international and Russian scientists and practitioners, Oslo Manual, internal regulatory corporations' framework. Methodological base of the research: system, process and logical scientific approaches to study processes of promising corporative types of intellectual capital generation and assessment aimed at subsequent transformation into digital innovations in the "open innovation" system.

Each stage of the research required general scientific and special methods to achieve proper efficiency, they were: analysis and synthesis, generalization and detailing, inductive and deductive, selective observation, grouping, formalization, analogy, abstract-logical, monographic, analytical, comparative, economic-statistical and economic-mathematical. Information sources were selected for the period of 1998-2022, taking into account the principles of complexity, relevance, transparency, materiality, comparability with the best international practices, consistency, preference, objectivity, reliability, relevance and up to date information .

J. A. Schumpeter [9] essentially connects initiation of innovations with generation of intellectual capital prerequisites. V. L. Inozemtsev [10] clearly names intellectual capital as opportunity for innovations, L.V. Yurieva et al. [11] highlight innovations as organizational intellectual capital element, V. P. Bagov et al. [12] consider it as an intellectual resource that predetermines organization’s ability to produce and sell its innovative products.

L. Edvinsson and M. S. Malone [13], B. Lev [14], A. Pulic [15] consider innovative component of organizational intellectual capital as a separate type of capital.

According to J.A. Schumpeter [9] determines practical implementation of scientific and technical initiatives and inventions as the essence of innovation, whereas underlining the meaning of an entrepreneur as a person of business amongst invention and innovation confirms importance of human capital.

Yu. V. Vertakova, E. S. Simonenko [16] point out that accomplishment of an invention or a discovery in a particular human activity area transforms underlying ideas of the invention into an innovation that contributes to new ideas emerge, eventually causing generation of new products (technologies). They associate invention with high-level innovation, defining products and technologies, social, economic, environmental and managerial processes as the subject of change. In the result there is an innovative spiral cycle: “scientific and technological progress - idea - innovation - scientific and technological progress - idea - innovation - ...”

In the Center for Economic and Social Research of the Republic of Tatarstan [17] new phenomenon, discovery, idea, method, etc., presented as a result of research, development / empirical work on increasing particular sphere of business efficiency is called innovation; its implementation creates innovation taking into account condition of human involvement that changes along with development of socio-cultural systems and regions. Innovation is shaped based on investment into new equipment, technologies, regulation

systems, labour organization and etc., that together shape intellectual organizational capital. In general, organizational capital of corporation is assessed according to the principle “inventions available in the patent portfolio mean future innovations”.

V.V. Platonov [18] recognizes innovation capital (industrial property, technological know-how) as a part of organizations’ intellectual capital along with the network, and human and organizational capitals.

B. Lev's model [14] “Value Chain Blueprint” describes intellectual capital on the basis of corporation's determination for innovations and integrates 9 parameters: internal renewal availability (research, personnel development, organizational processes); integrated acquired abilities (technologies, investment business); intellectual property; technological feasibility for innovations (clinical tests, approvals, prototypes); business networks (alliances, integrated corporations, customers and suppliers associations); customers (marketing alliances, brand value, value, customer drain); business on the Internet (website traffic, online orders, alliances on the Internet); performance (sales, including licenses, profit, market share, new products); future growth (periods for bringing new products to the market, planned initiatives, increase in results, etc.).

Human capital is described along with traditional qualities of a person: knowledge, skills, qualifications, competence, and other characteristics correlated with personality as well as innovations capability.

“Social innovation” term is always put together with the concept of “intellectual capital”, which comes up [19] when strengthening human factor via creation and introduction of systems for modernized personnel policy, professional retraining and growth of employees, social and professional adaptation of newly hired, bonuses and assessment of work results. It should be emphasized that K. Marx [20] when characterizing the economic category "capital" underlined its social aspect.

TABLE I. INFLUENCE OF FACTORS ON INTELLECTUAL CAPITAL FOR EVERY THREE PAST YEARS, ACCORDING TO FORM 4 - INNOVATION - ADAPTED FOR ORGANIZATIONS IN STAVROPOL REGION, UNITS

Factors	Factors' impact				
	1 – unessential	2 – essential	3 – crucial	4 – No clear answer	5 – Not applicable
<b>Human capital</b>					
Lack of qualified personnel	137	105	22	185	204
<b>Organizational capital</b>					
Uncertain economic benefits from intellectual property utilization	94	95	27	247	190
<b>Stakeholders' capital</b>					
Lack of own financial resources	66	181	94	157	155
Lack of State financial support	76	159	46	189	183
Lack of information on new technologies	150	70	22	197	214
Lack of information on sales markets	152	64	20	200	217
Underdeveloped cooperation networks	124	67	15	235	212
Underdeveloped innovations' infrastructure (intermediaries, information, legal, banking and other services)	143	68	25	237	180
Total	711	609	222	1215	1161

Source: calculation provided by the authors V.V. Manuylenko, G.A. Ermakova [19]

V.V. Manuylenko, A.A. Mishchenko [21] associate intellectual capital availability with personnel innovations emerge and regular improvement of erudition level with a professional team. Surely there is an opposite effect of intellectual capital sales and social or personnel innovations.

E.V. Petrukhina [22] recognizes following functions of organization's intellectual capital: education and future development of intellectual property; promotion of employees' innovative thinking, businessmen, scientists, and management teams that shape and test key models for reproduction within a single economic system or systems units.

Aggregated intellectual capital, as noted by V.V. Manuylenko, G.A. Ermakova [19], performs the function of corporations' innovative development. In terms of reproductive performance intellectual capital of corporations is simultaneously a prerequisite and result of innovations

creation process showing the result of previous or current innovative activities in the "open innovation" system. I.e., intellectual capital of corporations is a resource for future oriented innovative activities. Following the logics of the research, assessment of both intellectual capital and innovations in corporations is a challenging study issue.

#### A. Synergy of Intellectual Capital and Innovation Assessment

Considering available intellectual capital in corporations for future innovations, intellectual capital of corporations is assessed according to Form 4 – innovation (adapted). Still, in Form 4 – innovation "Information about organization's innovative activity" the following parts draw special attention: factors limiting innovation and organizational innovations.

Factors affecting each type of intellectual capital in organizations (human, organizational, stakeholders') – Table I.

TABLE II. ASSESSMENT OF POSSIBILITIES FOR CHANGES IN INTELLECTUAL CAPITAL DEVELOPMENT WITH ORGANIZATIONAL INNOVATIONS UTILIZATION IN CORPORATIONS OF STAVROPOL REGION, UNITS

Indicators	YY								
	2010	2011	2012	2013	2014	2015	2016	2017	2018
Number of corporations implementing organizational innovations during past three years	7	4	9	8	6	4	2	3	8
Number of corporations with no organizational innovations	435	459	455	449	438	450	690	650	811
<b>Organizational innovations / Human capital</b>									
Innovations concerning shift working hours	4	3	2	4	2	2	2	1	1
Introduction of corporative knowledge management systems	4	2	3	1	0	2	0	0	2
Personnel development means (corporate and / or individual trainings, creation / development of personnel training and advanced training institutions)	6	4	7	8	6	4	2	3	7
Introduction of new methods for employees motivation	0	0	0	0	0	0	1	0	1
<b>Organizational innovations / Organizational capital</b>									
Development and implementation of new or significantly corrected corporate (shareholders') strategy	5	3	4	5	4	4	2	2	2
Introduction of modern (IT-based) management methods in corporations	6	4	5	5	2	2	2	3	4
Development and introduction of new or significantly updated organizational structures in corporations	7	4	4	5	4	4	2	2	2
Application of modern systems for quality control, certification of goods, works and services	6	3	3	5	4	4	1	2	5
Creation of units specializing in research and development, implementation of scientific and technical achievements (technology and engineering centers, small innovation corporations)	3	2	1	3	2	2	2	1	2
Transferring certain functions and business processes to specializing contractors (outsourcing)	3	2	2	3	4	4	2	1	2
<b>Organizational innovations / Stakeholders' capital</b>									
Introduction of modern logistics and supply systems for raw materials, materials, components ("just in time", etc.)	4	3	3	3	2	1	2	1	4
New forms of strategic alliances, partnerships and other types of cooperation networks with product consumers, suppliers, Russian and foreign manufacturers	2	1	1	3	2	2	2	1	1

Source: calculation provided by the authors V.V. Manuylenko, G.A. Ermakova [19] as per data provided by Federal State Statistics Service for the North Caucasus Federal District [23]



Among the factors affecting intellectual human capital the following should be noted: lack of qualified workers, intellectual organizational capital, i.e. uncertain economic benefits from intellectual property utilization, intellectual stakeholder capital, i.e. factors associated with generation of financial resources: lack of own funds and/or State financial assistance; informational factors: lack of information about new technologies and sales markets; at the same time, such factors as underdeveloped cooperation networks and innovation infrastructure (intermediary, information, legal, banking and other services) should be also taken into account. Among significant financial nature factors that affect intellectual capital of organizations there is a lack of own funds and/or State financial assistance, which is followed by factors of underdeveloped innovation infrastructure, insufficient information about new technologies and sales markets as well as underdeveloped cooperation networks. No clear response was noted for such factors that affect intellectual capital of organizations as uncertain economic benefits from intellectual property utilization, underdeveloped innovation infrastructure and cooperation networks [19]. In general, when determining influence of factors on intellectual (human, organizational, especially stakeholders') capital, first prevail those corporations that have no clear response or with no applicable factors, then there are those that note insignificant, significant and crucial influence.

It should be noted that Form 4 – innovation – combines factors of efficient reproduction of intellectual capital (scientific and technological progress, the level of IT literacy, innovation and intellectual property policies, modern market infrastructure, institutional environment, etc.).

Alternative area for human, organizational and stakeholders' capitals development is organizational innovations. Merged application of intellectual capital and organizational innovation causes overturn of the principle: "intellectual capital existing in the organization means future innovations", ultimately creates opposite effect in the "open innovation" system. Intellectual capital in corporations influences creation and implementation of organizational innovations, which in turn affects intellectual capital, changing the demands, as well as its content (ratio between human, organizational and stakeholders' capital).

In the corporations of Stavropol Region, organizational innovations affect generation of human, organizational and stakeholders' capitals – Table II.

Organizational innovations mainly develop towards improvement of human capital in the area of personnel development (corporate and / or individual training, creation / development of personnel training and advanced training institutions). These organizational innovations serve to minimize risks with personnel insufficient scientific qualifications. From the standpoint of intellectual capital future development it is not good that in the region number of corporations with no organizational innovations prevails (435 - 811 units) over the number of corporations that have ones (2 - 9 units). Increase of corporations that do not carry out organizational innovations makes it difficult to reveal synergistic effect on the principle "intellectual capital

available in the corporation means future innovations" in the "open innovation" system.

In developing countries above human capital, resources and networks Oslo Manual [24] recognizes priority of information and communication technologies, their ownership and application by organizations. By the level of innovative development innovations are classified according to the criterion of technological parameters that takes into account essence of innovations, i.e. complex attraction of new digital, information and communication technologies. In the current environment, information and digital technologies become a priority issue demanding identification of characteristics for human, informational and digital capitals. According to KPMG research, factors hampering introduction of innovative technologies in Russia as well as in the world are insufficient maturity of processes / low automation, low level of IT-literacy of employees. Among the threats for digitalization Russian and world experts recognize the risks of information security, etc. [25-27].

### *B. Characteristics and Assessment of Corporations'*

#### *Intellectual Information Capital: Alternative Point of View*

V.A. Medvedev [28] puts together informational capital and level of control over information. V.L. Inozemtsev [10] designates databases as part of structural capital elements. B.B. Leontiev [29], L. Edvinsson and M.S. Malone [13] recognize information systems as constituent parts of organizational capital, and B.B. Leontiev [29] adds up accumulated knowledge bases.

In developing countries, information and communication technologies, their ownership and application by organizations are predominantly distinguished. In 1972, K.D. Arrow [30] identified link between special economic behaviour of intellectual resources and information creating and handling processes.

In existing environment, encoded and materialized information is represented in human capital. According to P.F. Drucker [31], knowledge is "information" that changes something or somebody or else is the cause of action that provides opportunities for different and more effective actions.

Corporations' financial managers should provide access to sources of knowledge and information, which are subsequently transformed into separate production resources for specific use. Constant monitoring of information relevance and demand serves its transformation into corporations' valuable resource, which is ensured only with creation of developed information and communications infrastructure, assumed that only up-to-date information processing technologies and information exchange are applied. IT-specialists categorized as intellectual workers are in demand.

Creation and development of corporations' intellectual capital is subjective to correctly built information exchange system. Paragraph 23 of Oslo Manual [24] pays special attention to information needs of analysts and politicians when they accumulate information about innovation activities or determine set of indicators. According to the golden rule of information mobilization, information exceeding value or its collecting costs makes its accumulation reasonable. At the

same time, it is important to take into account distinctive properties of corporations' intellectual capital, i.e. information asymmetry and liquidity. Processing of information results in knowledge that reflects human capital - when put together it represents powerful competitive tool. Science being one of the constituent parts of intellectual activity produces new knowledge, that is, new information; in order to transform it into knowledge it is advisable to present it and combine it with existing knowledge, establishing value of received information, attributing it to structure (superstructure), and providing guidelines for its application. Thus, knowledge existing in time interval  $t + 1$  is a complex functional relation of received information and human knowledge in corporations. As a result, importance of information in creation and development of intellectual capital in corporations, which should meet uniqueness requirement, is confirmed. The link between intellectual information and human capital of corporations becomes evident. In terms of liquidity degree information systems usually have medium liquidity, whereas databases and accumulated knowledge bases are low liquid.

As a result, it is fair to set aside intellectual information capital of corporations intended to ensure information security of intellectual activity; its functioning reveals risks of insufficiency and inconsistency of information capital in the existing environment, etc.

*C. Characteristics of Intellectual Digital Capital of Corporations, its Interaction with Human and Informational Capital*

Considering that corporations' "intellectual capital" concept is constantly changing over time, and accelerated introduction of digital technologies in economy and social sphere is one of the national development goals [32], intellectual digital capital is of great importance for Russian corporations. Information and digital technologies can ensure operation of corporations 24/7/365. Digitalization represents integration of digital technologies in order to improve project's performance by means of key processes adjustment [33]. Modern digital technologies are focused on transforming traditional business models and business processes in already existing industries. Innovations grounded in digital technologies create new electronic financial products and services and bring up to date their forms.

Technical financial industry having technological foundation shall incorporate provisionally basic end-to-end technologies: telecommunications, Big Data technologies, Internet of Things, industrial and simulated intellect, as well as interdisciplinary technologies: neuro-technologies, distributed ledger systems, elements of robotics and sensor technologies, quantum, new production technologies, technologies of virtual and augmented reality.

Hence, for most corporations, efficient digitalization process indicator represents "digital versions" of already existing conventionally traditional solutions, optimization results for actual business processes, whereas implementation of innovative changes in business models indicates future prospects. Corporations with limited digital technologies may provide basic services, such as balance sheet regulation, in the

"open innovation" system – Table III.

There are criteria for assessing satisfaction level of digital customers: regular compliance of requirements to provided services, timely receipt of information about digital products and services, assessment of the ratio between results and resources spent on digital products and services purchase, risk level at digital products or services receipt, overall satisfaction with digital products and services, digital customer feedback. Creation and development of digital stakeholders' capital in corporations should be aimed at exceeding stakeholders' expectations. It is true that digital stakeholders' capital in its essence represents digital asset value, expressed in relations with stakeholders, i.e. digital clients, which when efficiently managed by a corporation on the basis of special marketing activities maximizes value and ensures competitiveness, and in the quantitative aspect justifies amount of discounted cash flows obtained from real and potential stakeholders, i.e. digital clients, and excludes mobilization costs, that is lifetime value of real and potential stakeholders.

TABLE III. BASIC DIGITALIZATION AREAS FOR CORPORATIONS IN THE "OPEN INNOVATION" SYSTEM

Areas	Digital format functions	Financial technologies
Regulation and assessment of assets and liabilities	Assets and liabilities regulation in digital format	Big Data, block chain
Financial consulting	Financial consulting in digital format	Big Data, artificial intellect
Management Accounting	Online accounting	Cloud technologies
	Online reporting	SaaS
	System solutions	Cloud technologies, block chain
Infrastructure solutions	Identification solutions	block chain, artificial intellect, Big Data, SaaS
	distributed ledger and automated (smart) agreements	artificial intellect, block chain
	Information Security	computer education, artificial intellect, predictive analytics, block chain

Source: information provided by the author V. V. Manuylenko

One of the key corporation's competitive advantages is highly qualified personnel, namely central competence of corporation (capability to digital training and personnel development, infrastructure for long-term future growth) – employees with their own experience and qualifications, as well as unique abilities to digitalize economic processes. According American economist E. Denison study in 1929-1982, American economy growth occurred by 32% due to new labour force mobilization, 1.4% - due to increase of education level, 28% - progress in knowledge; 19% – new investment, 17% – modernization of production structure and labour organization [34]. Top three new key competencies of future demand include: flexible and critical thinking, creativity, emotional intelligence, which distinguish person from machine. Most demanded specialists for implementation of transformation programs are business analysts in process optimization and data analysis. Digitalization and innovation professional personnel should include IT specialists with different background (from data analytics, robotics and

interface design up to cyber security and integration), IT auditors with knowledge and skills in obtaining and analyzing digital information technology data [35], top risk management, internal audit, compliance, experts from other technological organizations that include various types of economic activities and research institutions. Creation, assessment of intellectual digital capital in the “open innovation” system should include professional judgment of specialists.

It is clear that initially digital intellectual capital causes risks of staff deficiency, insufficient level of scientific personnel qualifications and low scientific specialization. At the same time, introduction of financial technologies raises up threats of digital fraud in financial sector, leading to information security risks that also reveals links between intellectual, informational and digital capitals. Risks complementary to corporations’ intellectual digital capital functions are defined as possibility of wrong digital decision making. Synergy effect of merged intellectual, human, informational and digital capitals functioning should ensure development of digital innovations. Hence, digital innovations represent result of human, informational and digital capitals interaction, implementation of which results in relations and links rising up between corporations and digital clients in the “open innovation” system, that represents intellectual digital stakeholders’ capital.

### III. RESULTS

#### A. In Theoretical Block of the Research

- determined that innovative activity in the “open innovation” system derives from corporations’ intellectual capital elements;
- all types of innovations are based on technologies, among which - in the context of digital economy development - informational and digital technologies draw special attention; that makes necessary to research intellectual, information and digital capitals, i.e. basis of digital innovations development in the “open innovation” system;
- established connection between the concepts of organizations’ “intellectual capital” and “social”, “personnel” innovations, which mainly is reflected in strengthened human factor by means of development and introduction of modernized personnel policy systems, regular improvement of personnel / team education level, etc.;
- identified corporations’ intellectual informational capital function in the "open innovation" system, i.e. provision of information security for intellectual activity;
- described supplementary risks of corporations’ intellectual digital capital functioning at initial stage (staff deficiency, personnel scientific qualifications level, low scientific specialization) and subsequent threats caused by digital fraud possibilities in financial sector, leading to information security risks in the system "open innovation".

#### B. In Practical Block of the Research

- considering that available intellectual capital in corporations represents future innovations, Form 4 – innovation – was adapted to corporations’ intellectual capital assessment in the following areas: factors limiting innovations, organizational innovations;
- found that development of intellectual capital in most corporations is complicated by insufficient consideration of affecting factors (corporations with no answer or no factor prevail, organizational capital is affected by uncertainty of economic benefits from intellectual property utilization) that makes necessary call for professional judgment in assessment process in order to eliminate controversial interpretation of results, errors and manipulations, as well as decrease of corporations that carry out organizational innovations, whereas their influence is noted on building up human, organizational and stakeholders’ capital in corporations of Stavropol Region;
- based on corporations’ intellectual capital assessment according to adapted Form 4 – innovation, its key advantages and disadvantages are identified – Table IV;
- indicated sequence of intellectual capital assessment stages based on adapted Form 4 – innovation, as follows: 1) determination of the factors influencing intellectual capital, 2) assessment of changes possibility in intellectual capital development via organizational innovations, 3) identification of the key advantages / disadvantages of intellectual capital assessment by Form 4 – innovation, 4) generation of motivated professional judgment of specialists;
- found that most corporations consider digitalization processes efficiency indicator in the “open innovation” system as “digital versions” of conventionally traditional existing solutions, results of actual business processes optimization, whereas innovative changes in business models are taken for future projections.

TABLE IV. KEY ADVANTAGES AND DISADVANTAGES OF CORPORATIONS’ INTELLECTUAL CAPITAL ASSESSMENT ACCORDING TO ADAPTED FORM 4 – INNOVATION

Advantages	Disadvantages
Development of scientific level with actually obtained results of intellectual activity.	Lack of internal regulatory framework governing intellectual capital assessment process.
Development of inventive level represents a fact of created patented inventions, and good perspectives for inventing object	Based on non-formalized methods, expert assessments method with subjective nature, which do not represent importance of intellectual capital in added economic value and its impact on the market value.
Development of design level is demonstrated in patented industrial samples and applications of industrial design and exterior of commercial product submitted for patenting.	Tendency to collection of quantitative indicators, and their application without strict analytical dependencies.
Software creation represents	No mechanism for transfer of ideas

object for development of officially registered computer programs and / databases and / purchased applications for the official registration of computer programs and / databases	and high technologies into market products / or intellectual potential into capital
	No mechanism for transforming human and stakeholders' capital into capitalized assets, since market qualifications are recognized via certificates and licenses, reputation – via brands, knowledge – via intellectual property rights.
	No demo effect of ultimate goal from intellectual capital application that breaks main property of capital – its ability to generate income.
	Lack of tools for assessing and regulating risks associated with intellectual capital the functioning.

Source: research provided by the authors V. V. Manuylenko, G. A. Ermakova [19]

### C. In Methodological Block of the Research

- established link between human and informational capital, when knowledge in a time period  $t + 1$  represents complex function of received information and knowledge of a subject of business – i.e. a person in a corporation;
- determined that informational capital in the “open innovation” system reflects quality of information systems, databases, incorporating efficiency of information application, administrative systems and organizational structures, information products and technologies, etc., eventually performing a function of ensuring information security for intellectual activity;
- proposed that quality of information systems is assessed by means of satisfaction level of internal clients – employees of corporations, taking into account main theoretical and methodological provisions of the assessment – Fig. 1. The following are priorities for modernization projects: automation of information system, improvement of transparency and quality of information, availability of up-to-date information, regularity, efficiency and timeliness of information receipt, importance of regional aspect in information systems operation in terms of global trends. Assessment of satisfaction level with information systems by internal clients allows identify at relatively low costs wide range of problem areas, and represents efficient tool for development of informational capital. It is important that corporative cultures accept informational systems expressing affiliation of informational and cultural capitals;
- merged performance of intellectual, human, informational and digital capitals ensures development of digital innovations, causing synergy effect in the “open innovation” system. In digital economy environment, particularly these types of corporations' intellectual capital penetrating into each other create intellectual wealth, and obtain strategic national importance;
- outlined main areas of digitalization in traditional corporations in the context of functions revealed in digital format and financial technologies in the "open

innovation" system (management accounting - system solutions - cloud technologies, block chains, API; infrastructure solutions - information security - computer learning, artificial intelligence, predictive analytics, block chains, etc.);

- determined that at the initial stage digital intellectual capital is accompanied by risks of lack of personnel, low level of personnel scientific qualifications, low scientific specialization of personnel; subsequently introduction of financial technologies causes threats of developing digital fraud in financial sector, leading to information security risks in “open innovation” system.
- found that synergy of intellectual, informational, digital and human capitals in corporations contributes to development of digital innovations in the “open innovation” system.

### IV. DISCUSSION

The authors believe that developed toolkit for creation and assessment of intellectual, informational and digital capitals aimed at digital innovations development requires steady upgrade of information and digital technologies in the “open innovation” system. I. Kukhnin, specialist of Deloitte CIS Research Center [36], evaluated technologies in terms of their potentials as follows: artificial intelligence - 0.88, computer learning - 0.84, predictive analytics – 0.82, deep learning - 0.79, and Big Data - 0.78, block chain - 0.54, augmented reality - 0.58 and virtual reality – 0.55. The best effect of digital technologies application was brought by solutions on Big Data and predictive analytics - 40%, robotization - 38%. Big Data and predictive analytics are used in the following key areas - operational intelligence and analytics - 52%, customer service, sales and marketing - 35% [25-27].

It is important to take into account that information and digital technologies development has certain specifics in corporations in terms of their different business activity types. Analysis of cross sector integration of industrial CIOs from more than 90 countries, conducted by Gartner [37], shows the following ranking of economic activities according to potential impact of digital transformation: media – 30%, banking and investment services – 26%, telecommunications – 25%, insurance and transport – 22%, services – 19%, government – 18%, healthcare – 17%, retail – 15%, housing and utilities, industry – 14%, healthcare providers – 13%, education – 12%, wholesale trade – 11%, natural resources – 7%.

Corporations' digital innovations development implies:

- identification of logical links between the terms: “intellectual capital” and “social”, “personnel” innovations;
- identification of informational capital assessment stages;
- identification of risks linked with intellectual capital creation and application.
- Corporations' digital innovation development is possible in the following areas:

- corporations' digital strategy;
- intellectual capital development trends; separate research of intellectual stakeholders' capital of corporations was carried out by S.S. Galazova et al. [38];
- optimal combination of human, informational and digital capitals: E.A. Posnaya et al. [39], T.V. Shabunina et al. [40]; E.V. Rodionova et al. [41] – regional level;
- complex intellectual capital reproduction: E. Zhilenkova et al. [42];
- intellectual digital capital creation risk management: V.A. Kunin and D. Mikhailovsky [43].

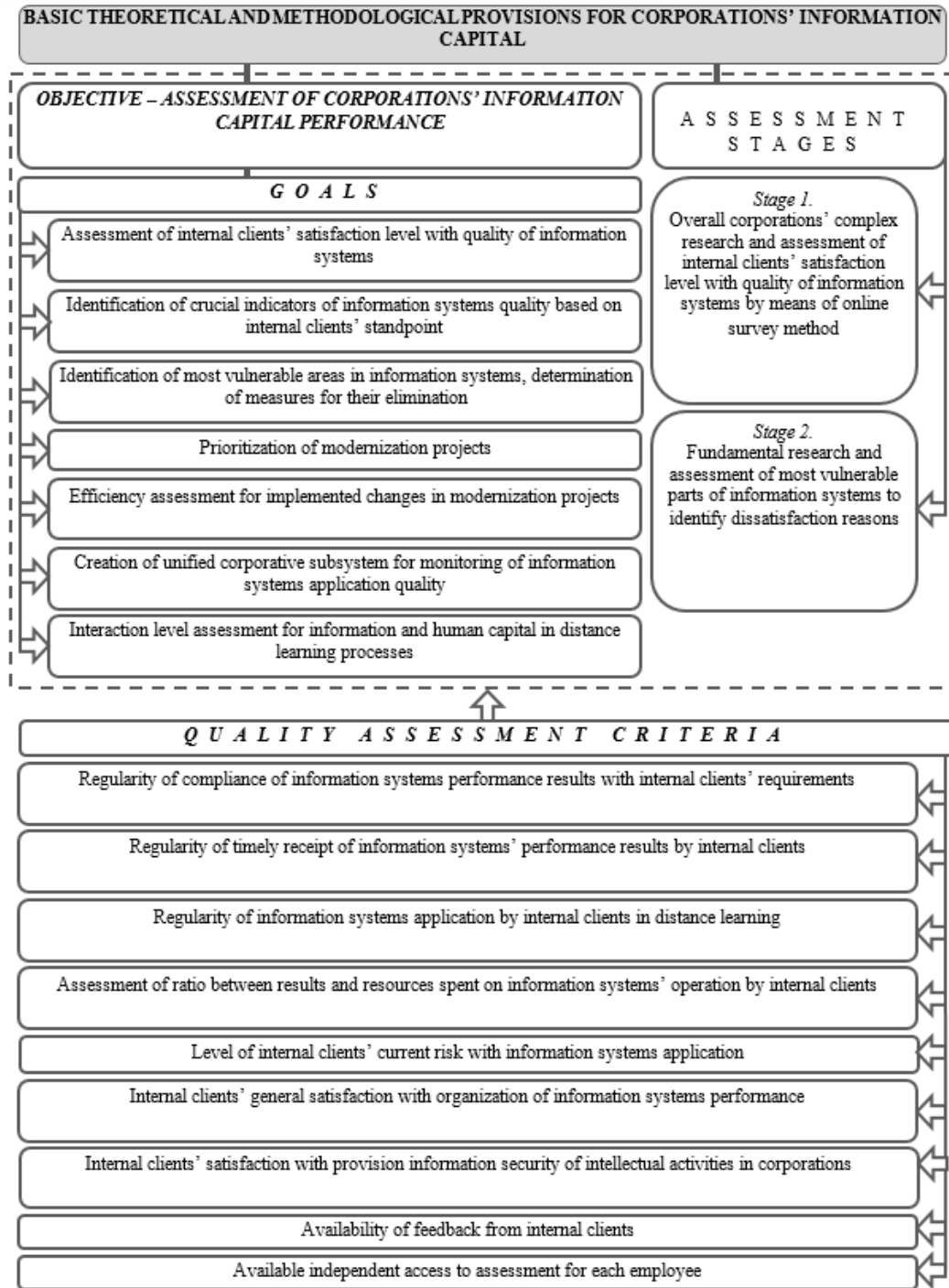


Fig. 1. Logical and Structural Chart for Creation and Assessment of Informational Capital in Corporations (Developed by Authors V. V. Manuylenko, G. A. Ermakova)

## V. CONCLUSIONS

It is important to note that development of digital innovations predominately on intellectual, informational and digital capitals foundation in the “open innovation” system serves to reduce research and development costs, create potential for efficiency improving as well as potential synergy between informational and digital innovations. Thus, conception was created and implemented as follows: intellectual informational and digital capitals functioning in corporations simultaneously with human and stakeholders’ capital represent future potential digital innovations in the “open innovation” system. Research possibilities are limited with underdeveloped legislation that regulates digital economics processes, which predetermine development of digital innovations in corporations. Different corporations may independently formulate guidelines for digital innovations development based on proposed scientifically grounded toolkit.

## REFERENCES

- [1] F. Schiavone, D. Leone, A. Caporuscio, and A. Kumar, “Revealing the role of intellectual capital in digitalized health networks. A meso-level analysis for building and monitoring a KPI dashboard”, *Technological Forecasting and Social Change*, vol. 175, 121325, 2022. <https://doi.org/10.1016/j.techfore.2021.121325>
- [2] K. S. Al-Omouh, D. Palacios-Marqués, and K. Ulrich, “The impact of intellectual capital on supply chain agility and collaborative knowledge creation in responding to unprecedented pandemic crises”, *Technological Forecasting and Social Change*, vol. 178, 121603, 2022. <https://doi.org/10.1016/j.techfore.2022.121603>
- [3] S. Kusi-Sarpong, M. S. Mubarak, S. A. Khan, S. Brown, and M. F. Mubarak, “Intellectual capital, blockchain-driven supply chain and sustainable production: Role of supply chain mapping”, *Technological Forecasting and Social Change*, vol. 175, 121331, 2022. <https://doi.org/10.1016/j.techfore.2021.121331>
- [4] X. Wang, R. Sadiq, T. M. Khan, and R. Wang, “Industry 4.0 and intellectual capital in the age of FinTech”, *Technological Forecasting and Social Change*, vol. 166, 120598, 2021. <https://doi.org/10.1016/j.techfore.2021.120598>
- [5] E. García-Meca, and I. Martínez, “The use of intellectual capital information in investment decisions. An empirical study using analyst reports”, *International Journal of Accounting*, vol. 42, no. 1, pp. 57–81, 2007. <https://doi.org/10.1016/j.intacc.2006.12.003>
- [6] G. Secundo, V. Ndou, P. D. Vecchio, and G. De Pascale, “Sustainable development, intellectual capital and technology policies: A structured literature review and future research agenda”, *Technological Forecasting and Social Change*, vol. 153, 119917, 2020. <https://doi.org/10.1016/j.techfore.2020.119917>
- [7] A. Kalkan, Ö. Ç. Bozkurt, and M. Arman, “The impacts of intellectual capital, innovation and organizational strategy on firm performance”, *Procedia – Social and Behavioral Sciences*, vol. 150, pp. 700–707, 2014. <https://doi.org/10.1016/j.sbspro.2014.09.025>
- [8] F. Ricci, V. Scafarto, S. Ferri, and A. Tron, “Value relevance of digitalization: The moderating role of corporate sustainability. An empirical study of Italian listed companies”, *Journal of Cleaner Production*, vol. 276, 123282, 2020. <https://doi.org/10.1016/j.jclepro.2020.123282>
- [9] J. A. Schumpeter, *Economic development theory. Capitalism, socialism and democracy*. Moscow: EKSMO, 2007.
- [10] V. L. Inozemtsev, *Outside economic society: postindustrial theories and post-economic trends in the modern world*. Moscow: Academia-Science, 1998.
- [11] L. V. Yurieva, O. V. Bazhenov, and M. A. Kazakova, *Integrated management accounting and analysis of innovative activities in metallurgical holdings*. Moscow: INFRA-M, 2013.
- [12] V. P. Bagov, E. N. Seleznev, and V. S. Stupakov, *Intellectual capital management: Textbook*. Moscow: Publishing House “Cameron”, 2006.
- [13] L. Edvinsson, and M. S. Malone, *Intellectual capital: Realizing your company's true value by finding its hidden brainpower*. New York, NY: Harper Business, 1997.
- [14] B. Lev, *Intangibles: Management, measurement, and reporting*. Washington, DC: Brookings Institute Press, 2001.
- [15] A. Pulic, “Intellectual capital – Does it create or destroy value?” *Measuring Business Excellence*, vol. 8, no. 1, pp. 62–68, 2004. <http://dx.doi.org/10.1108/13683040410524757>
- [16] Yu. V. Vertakova, and Ye. S. Simonenko, *Innovation management: Theory and practice: Workbook*. Moscow: Eksmo, 2008.
- [17] Center for Economic and Social Research of the Republic of Tatarstan under the Cabinet of Ministers of the Republic of Tatarstan, *Methodological recommendations for innovative activity monitoring in the Republic of Tatarstan*. 2007. [Online]. Available: <https://doc4web.ru/raznoe/metodicheskie-rekomendacii-po-monitoringu-innovacionnoy-deyateln.html>
- [18] V. V. Platonov, *Intellectual capital: Assessment and management*. St. Petersburg: Publishing house of SPbGUEF, 2012.
- [19] V. V. Manuylenko, and G. A. Ermakova, *Assessment of intellectual capital in Russian corporations: Monograph*. Moscow: Prospect, 2020.
- [20] K. Marx, *Capital. Criticism of political economy*. Vol. 1. Moscow: Publish house for political literature, 1969.
- [21] V. V. Manuylenko, and A. A. Mishchenko, “Evaluation of intellectual capital as strategic factor in development of innovations in commercial organizations”, *Financial Analytics: Problems and Solutions [Finansovaya Analitika: Problemy i Resheniya]*, no. 39(321), pp. 16–27, 2016.
- [22] E. V. Petrukhina, “Role of intellectual capital in ensuring innovative development of enterprises”, in *Innovative and creative economy development issues: almanac of reports of international scientific and practical conference*, N. A. Gorelov, O. N. Melnikov, and E. G. Abramov, Eds. Moscow: Creative Economy, 2010, pp. 356-360.
- [23] Territorial body of the Federal State Statistics Service for the North Caucasus Federal District, *Official website*. [Online]. Available: <https://stavstat.gks.ru/>
- [24] OECD, Eurostat, *Oslo Manual. Recommendations on innovations data collection and analysis*, 3rd ed. Moscow: CISN, 2010.
- [25] KPMG, *Banking fraud: Are Russian banks ready for the challenge? 2019*. [Online]. Available: <https://assets.kpmg/content/dam/kpmg/ru/pdf/2019/12/ru-ru-global-banking-fraud-survey.pdf>
- [26] KPMG, *Digital technologies in Russian companies*. 2019. Available: <https://assets.kpmg/content/dam/kpmg/ru/pdf/2019/01/ru-ru-digital-technologies-in-russian-companies.pdf>
- [27] KPMG, Institute of Internal Auditors, *Research of the current situation and development trends of internal audit of financial organizations in Russia in 2018*. 2019. [Online]. Available: <https://www.iiar.ru/contact/ru-ru-internal-audit-in-financial-institutions-and-companies-report.pdf>
- [28] B. A. Medvedev, *Facing challenges of post-industrialism: Glimpse back, present and future of Russia*. Moscow: Alpina Publisher, 2003.
- [29] B. B. Leontiev, *The price of intelligence. Intellectual capital in the Russian business*. Moscow: Publishing center “Aktioner”, 2002.
- [30] K. D. Arrow, “Information as an economic commodity”, *HSE Journal of Economics [Ekonomicheskij zhurnal VShE]*, vol. 16, no. 2, pp. 161–171, 2012.
- [31] P. F. Drucker, “Knowledge worker productivity: The biggest challenge”, *California Management Review*, vol. 41, no. 2, pp. 79-94, 1999.
- [32] President of the Russian Federation, *Decree of the President of the Russian Federation of May 7, 2018 No. 204 “On national goals and strategic development tasks of the Russian Federation for the period up to 2024”*. 2018. [Online]. Available: <http://www.kremlin.ru/acts/bank/43027>
- [33] National Research University. Higher School of Economics. *Development Center, Innovative financial technologies and services market*. 2019. [Online]. Available:



- <https://dcenter.hse.ru/data/2019/12/09/1523584041/%D0%A0%D1%8B%D0%BD%D0%BE%D0%BA%20%D1%84%D0%B8%D0%BD%D0%B0%D0%BD%D1%81%D0%BE%D0%B2%D1%8B%D1%85%20%D1%82%D0%B5%D1%85%D0%BD%D0%BE%D0%BB%D0%BE%D0%B3%D0%B8%D0%B9-2019.pdf>
- [34] Uspeshnyy menedzhment, Economic growth forecasting models and methods. [Online]. Available: <http://www.mansuccessful.ru/lavs-313-2.html>
- [35] KPMG. Official website. [Online]. Available: <https://home.kpmg/ru/ru/home.html>
- [36] Deloitte CIS Research Center, Certain financial technologies as a tool for sustainable business development in Russia and Kazakhstan. Financial technology market trends. 2018. [Online]. Available: <https://www2.deloitte.com/kz/ru/pages/research-center/articles/chastnye-finansovye-tehnologii-kak-instrument-ustojchivogo-razvitiya-biznesa-rossii-kazahstane.html>
- [37] Gartner, Official website. [Online]. Available: <https://www.gartner.com>
- [38] S. S. Galazova, V. V. Manuylenko, B. T. Morgoev, and N. V. Lipchiu, "Formation of stakeholders' client capital of trade institutions", *European Research Studies Journal*, vol. 20, no. 4B, pp. 398-411, 2017. <http://dx.doi.org/10.35808/ersj/898>
- [39] E. A. Posnaya, I. G. Vorobyova, E. M. Sokolova, and M. P. Leonova, "The role of human factors in the bank capital evaluation framework", *European Research Studies Journal*, vol. 20, no. 1, pp. 148-154, 2017. <https://doi.org/10.35808/ersj/604>
- [40] T. V. Shabunina, S. P. Shchelkina, and D. G. Rodionov, "Regional habitat as a factor of the human capital assets development in Russian regions", *Journal of Social Sciences Research*, vol. 2018, no. S3, pp. 313-317, 2018. <https://doi.org/10.32861/jssr.spi3.313.317>
- [41] E. V. Rodionova, Z. Kuzminykh, and E. Gamova, "Regional differentiation of digital economy development in the Russian Federation", in *Proceedings of the 2019 International SPBPU Scientific Conference on Innovations in Digital Economy (SPBPU IDE '19)*, October 24-25, 2019, St. Petersburg, Russia. Association for Computing Machinery, 2019. <http://dx.doi.org/10.1145/3372177.3374656>
- [42] E. Zhilenkova, M. Budanova, N. Bulkhov, and D. Rodionov, "Reproduction of intellectual capital in innovative-digital economy environment", *IOP Conference Series: Materials Science and Engineering*, vol. 497, 012065, 2019. <https://doi.org/10.1088/1757-899X/497/1/012065>
- [43] V. A. Kunin, and D. Mikhailovsky, "The concept of end-to-end risk management of commercial business organizations under conditions of digitalization of Russian economy", in *Advances in Economics, Business and Management Research: Proceedings of the III International Scientific and Practical Conference "Digital Economy and Finances" (ISPC-DEF 2020)*. Atlantis Press, 2020, pp. 12-16. <https://dx.doi.org/10.2991/aebmr.k.200423.003>

# An Algorithm for Providing Adaptive Behavior to Humanoid Robot in Oral Assessment

Dalia Khairy<sup>1</sup>, Salem Alkhalaf<sup>2</sup>, M. F. Areed<sup>3</sup>, Mohamed A. Amasha<sup>4</sup>, Rania A. Abougalala<sup>5</sup>

Department of Computer Teacher Preparation, Damietta University, Damietta, Egypt<sup>1,4,5</sup>

Department of Computer Science, Qassim University, Alrass, Saudi Arabia<sup>2</sup>

Department of Computer Science, Damietta University, Damietta, Egypt<sup>3</sup>

**Abstract**—Assistance humanoid robots (AHR) are the category of robotics used to offer social interaction to humans. In higher education, the teaching staff supports the acceptance of AHRs as a social assistance tool during the learning activities, with the whole responsibility of the correct operation of the device and providing a more comprehensive view of the objectives and significance of AHR use. On the other hand, students deal with AHRs either as a friend or control figures as a teacher. This paper presents an algorithm for AHRs in oral assessments. The proposed algorithm focuses on four characteristics: adaptive occurrence, friendly existence, persuasion, and external appearance. This paper integrates AHRs in higher education to improve the value of psychological and social communication during oral assessment where can assist students in dealing with challenges, such as shyness, dissatisfaction, hesitation, and confidence, better than a human teacher can. Thus, AHRs have increased students' self-confidence and enriches active learning.

**Keywords**—Algorithm; humanoid robot; social robots; oral assessment; assistance robots; higher education; adaptive behavior robot

## I. INTRODUCTION

Robots are becoming beneficial components of the educational ecosystem. With their different capacities, which range from the ability to see people and their environment to the ability to reason and explain circumstances and people's emotions, robots are becoming useful components of the educational ecosystem. These robots have a physical presence as well as multimodal interaction skills, which are equally crucial. With their human-like look, humanoid robots bring even another dimension to our understanding of social cues and body language: Keys to more intuitive and natural human-robot interaction and robots in education interacting [1].

As humanoid robots become more prevalent and people engage with them, the social intelligence of artificial agents is receiving attention. The important social skills presented by the standardized evaluation method for humans, Evaluation of Social Interaction (ESI), include approaches, speaking, turn-taking, gazing, and gesturing. When speaking to others, people employ co-speech gestures to accentuate their words, convey their intentions, or give detailed descriptions. Co-speech gestures have been shown to have strong impacts in numerous social science studies [2], and a neuroscience study supports motions are created by human professionals. The only movements that may be made are those that were considered during the design stage, even though hand-produced gestures

are natural and human-like. Furthermore, it takes significant human effort to create links between gestures and verbal phrases [4].

Due to their physical resemblance to humans, humanoid robots can give real-time feedback and interact with people more effectively. They have improved social abilities and are designed to show emotion through gestures, intonation, and facial expressions, as well as to respond with the right body language [5]. They can also display feelings, including shock, fear, rage, and disgust. Compared to human teachers, humanoid robots can assist in resolving issues connected to shyness, frustration, reluctance, and confidence. Humanoid robots are being widely employed in many nations, particularly for special education, and can assist students in dealing with challenges, such as shyness, dissatisfaction, hesitation, and confidence, better than a human teacher can. One of the factors contributing to humanoid robots' success in achieving learning objectives is that they never get tired, regardless of how many mistakes a pupil makes. Some humanoid robots support telepresence, which enables instructors to connect to the classroom remotely using display systems, which are typically built into the torsos of robots [6].

So, in the present paper, we focus on oral assessment to improve the value of psychological and social communications during oral assessment and to increase students' self-confidence and enrich active learning.

The remaining of this paper is organized as follows. Section 2 stated the problem statement. Section 3 is related to the work presented here. Section 4 presents an algorithm for AHRs. Section 5 describes the four recognized characteristics: adaptive occurrence, friendly existence, persuasion, and external appearance. These are described in Section 6. We conclude the paper and point future directions in Section 7.

## II. PROBLEM STATEMENT

Assistance humanoid robots (AHRs) in higher education have become an enriching teaching tool. They are different from portable devices such as smartphones, and tablets. AHRs have been described as human-looking forms with a head, arms, legs, and torso. Moreover, AHRs have characterized automation, repeatability, flexibility, digitization, anthropomorphism, body motion, and interaction [7]. Furthermore, AHRs encourage students to interact with reality in studying halls during the oral assessment process [8]. Consequently, AHRs have been characterized by personifying,

which provides more social interaction with students, such as instantly addressing them by name. In this environment, robots provide students with a more honest and realistic interaction than other technological devices [9].

On the other hand, dialogues support social interaction and create a shared experience of knowledge. Dialogue flow also provides a significant factor in increasing the quality of conversations [10]. As a result, a high degree of dialogue flow corresponds to favorable self-esteem, has an influence on the individual understanding of belonging, and can encourage solidarity [11].

Since dialogue flow can encourage social bonds and satisfy social conversation needs, education institutions should manage this knowledge when considering using AHRs in student oral assessments. Consequently, this can contribute to promoting dialogue between students and AHRs, especially in oral assessments. This paper has presented an algorithm for designing AHRs in higher education and put the main features of its external appearance and behavior.

### III. RELATED WORK

Artificial intelligence, sensors, mechatronics, and power are all components of humanoid robots. The main aim of modern humanoid robots is to acquire the ability to recognize visual expressions and perceptions to solve tasks, such as correctly predicting the emotion of a human by monitoring their visual facial expressions. Therefore, the only information that humanoid robots need supplied is the data that will supply enough relevant information to be processed, allowing them to do and expand the range of learning and performing activities that are already available to them. It will be up to the algorithms and other methodologies, such as deep learning and neural networks, to extract the features from photos that have been given to them. All these goals for humanoid robots present considerable difficulties and processing power needs, and it should be highlighted that a humanoid robot cannot employ this kind of enormous computing power alone. To install the cloud, which will further analyze the information and give it back to the humanoid robot, the humanoid robots must absorb, integrate, and collect the information from the environment [12].

Research in the disciplines of robotics that is currently considered critical technologies include multisensory perception, cognition, and man-machine interaction. Robotic systems are given theses and procedures from the field of artificial intelligence (AI). Significant biological principles can be used as recruiting tools and role models for robots. With remarkable accuracy, the human body is replicated in its greatest potential kinematic form. The main new area of research in AI is humanoid, complicated mechatronic systems inspired by biology. The psychological features of humanoid robots are just as fascinating as their technological ones [13].

More robots are being created for use in practical fields, such as education, healthcare, eldercare, and other assistive applications. For robots to have a positive impact on human life, there must be natural human-robot interaction (HRI). A human-like interaction is built on an understanding of the other person's needs and emotional condition at the time of the

interaction. To achieve this goal, Chiara and others proposed an ecological technology called thermal infrared imaging, which can give information on physiological characteristics related to the subject's emotional state. This ecological technology was presented and surveyed here. The technology can lay the foundation for the continued development of powerful social robots as well as for HRI. In the literature, thermal IR imaging has already proven effective for identifying emotions. This review can serve as a roadmap and encourage the usage of thermal IR imaging-based affective computing in HRI applications, which are meant to enable a natural HRI with a focus on people who find it challenging to convey their feelings [14].

Aburlasos et al. [15] proposed an organized and comprehensive modeling behavioral method to guide the activities of two interactive NAO robots with the goal of gradually evoking the Gestalt game in the mind of an autistic child. More specifically, the goal is to teach Gestalt's game first to robots and then to autistic children. In the end, children with autism will discover that playing with other children is more fun than playing alone with robots.

Teaching robots using hand gestures is now possible thanks to a framework developed by Mazhar et al. [16]. The background invariant robust hand gesture detector is the foundation of the suggested system. This was accomplished by applying a modern convolutional neural network that has already been trained, Inception V3, to the classification of 10 hand movements. The experiment validates the effectiveness of the suggested framework and guarantees a natural way to program robots. The combination of Kinect V2 and Open Pose allows the robot to understand its distance from the human worker to ensure the safety of nearby human coworkers.

Alemi, Meghdari, and Haeri [17] conducted a study in which they attempted to examine the attitudes of young EFL learners toward RALL. An experiment with a humanoid robot acting as a teacher's helper was carried out in a private kindergarten in Iran. They monitored the students' motivation, interaction, and anxiety as they studied with the humanoid robot. The outcomes demonstrated that the students' motivation increased because of their positive interactions with the humanoid robot. Additionally, they showed no indicators of nervousness while engaging with the humanoid robot during the learning process since it made the classroom feel welcoming. In conclusion, this study has served as a superb model for subsequent investigations using humanoid robots in second robots, which are seen as attractive and practical instruments in language learning and teaching circumstances. They can also accommodate the various needs of the students.

The field of social robotics is undergoing significant change because of neuroscience-based human-robot interaction, which is also improving our understanding of the human brain. Recent findings have demonstrated that more sophisticated analysis techniques and the trend of gathering data during real-time, embodied interactions with robots can deepen our understanding of the fundamental mechanisms underlying social cognition beyond simply perceiving robots in screen-based experiments. The development and design of the next generation of social robots, the same robots that may one

day function as social companions who support and care for their owners, stands to benefit from the additional (and natural) knowledge gained from this basic human brain research [18]. However, in less than 10 years, major problems about human-robot interaction have emerged from the field of neuroscience, such as: How can our relationships with these unique, mechanical companions benefit from the complex neural architecture of the human brain? How does the representation of social cognition evolve as robots become more pervasive in our social lives? Future research fusing human neurology and social robotics will shed light on how to live with autonomous robots that connect with us socially.

#### IV. HUMANOID ROBOT IMITATION LEARNING

Imitation learning refers to a humanoid robot's acquisition of skills or behaviors by observing a teacher demonstrating a given task. With motivation from neuroscience, imitation learning is a significant portion of AI, and human-computer interaction, and in turn, is taking a part in the future of robotics [19]. Another approach to imitation learning depends on trial and error, which introduces valuable models to understand desired behavior from a set of collected instances [20].

Imitation learning works by extracting features about the instructor's actions and the surrounding environment, including any manipulated objects, and understanding a mapping between the current position and revealed behavior. Traditional machine learning algorithms do not scale to high-dimensional agents with high degrees of freedom. Particular algorithms are therefore needed to build satisfactory representations and predictions to simulate motor processes in humans [21].

In the complete method of robotic imitation learning, demonstration teaching provides a teaching sample containing made options, operation characterization characterizes the options within the teaching sample as valid forms that the golem will acknowledge. The last word goal of imitative learning is to create the robot "master" behavior, which implies that the robot has got to reproduce the behavior and generalize the behavior into different unknown scenes. The method that robots use the teaching data to "master" behavior will be referred to as operation imitation. This thought methods of operational imitation are roughly divided into three categories:

activity biological research, inverse reinforcement learning, and adversarial imitation learning.

#### V. FROM HUMAN ASSESSMENT TO ROBOT

Although oral assessment has a wide record in examination practice in higher education and is a well-established component of such proceedings, concerns remain regarding its use. Therefore, there has been some move away from oral assessment for students, partly through a consideration of validity, trustworthiness, and justice [22].

Given the long history of oral assessment, which has some distinct advantages, it reflects the oral form document of communication that dominates professional practice. Also, it can test the limits of a student's knowledge and understanding. Moreover, it is thought to be an especially practical way of assessing individual types of capabilities. Furthermore, it supports intra-personal differences, such as trust and self-awareness [23].

AHRs have become an essential component of higher education infrastructure as teachers; they might take over assigned tasks that professors carry out. For example, AHRs deliver hints upon request to students [24].

In this case, the AHR keeps students' inspiration at a high level by prompting them. AHRs can be friends with students and share learning with them. Also, AHRs could assist in preserving student inspiration at a high level by encouraging them. Consequently, the AHR can play a valuable role in oral assessments. Oral assessment supports active learning through its empowerment of creating confidence among teaching staff and students during the oral assessment process of AHR technology [25]. That is, in turn, it may create more confidence for professors and higher education about the integration of AHR technology into their strategy.

#### VI. PROPOSED ALGORITHM

This paper presents an algorithm for AHRs, which is described by four recognized characteristics: adaptive occurrence, friendly existence, persuasion, and external appearance. This is detailed in this section.

##### Algorithm1: AHR generates gestures during dialogue acts information.

- Step 1: Start.
- Step 2: Collect speech datasets.
- Step 3: Determine the consideration of speech.
- Step 4: Define the dialogue acts based on speech consideration.
- Step 5: Classify the dialogue acts information into categories.
- Step 6: Discover the gestures happening combined with dialogue acts.
- Step 7: Extract main motion features.
- Step 8: Cluster these features to reduce dimensionality and increase the representability of gesture motions.
- Step 9: Develop conceptual models to connect dialogue information with gesture motion.
- Step 10: Analyze the intention of dialogue information related to gestures motion.
- Step 11: Discover the period of different gestures motion regarding dialogue information.
- Step 12: Determine the phases of gestures motion generation during dialogue information.
- Step 13: End.

Fig. 1. The AHR Generates Gestures during Dialogue Acts Information.

### A. Friendly Existence

Friendly existence expresses sociable presences. Students tend to humanize robots as technological devices, such as the AHR, and treat them socially when the technology shows social cues. The aim of this algorithm is to investigate the dialogue flow with AHR in the real oral assessment environment and to examine the extent to which oral flow is present in AHR–student interactions. Algorithm 1 described in Fig. 1 shows how AHR generates gestures. Dialogue acts as

information in oral assessments. This algorithm classifies dialogue gestures into categories and develops conceptual models to understand the intention of dialogue during oral assessment. Algorithm 1 can be compatible with generating hand gestures. It limits the whole arm's movements and manages the five fingers' movements. Also, there is a lack of head and torso motion. Another limitation can be indicated by eye gazing control issues. Algorithm 2, described in Fig. 2, generated laughing gestures during the dialogue assessment.

**Algorithm 2: AHR generates laughing gestures during dialogue information.**  
**Step 1:** Start.  
**Step 2:** Collect laughing speech dataset.  
**Step 3:** Analyze laughter types.  
(social, bitter, dumbfounded, and softening laughter).  
**Step 4:** Analyze laughter style.  
(secretly, giggle, guffaw, and sneer).  
**Step 5:** Analyze intensity level.  
**Step 6:** Analyze laughter function.  
(funny, amused, joy, mirthful laugh, social polite laugh, bitter/embarrassed laugh, self-conscious laugh, inviting laugh, contagious laugh, depreciatory/derision laugh, dumbfounded laugh, untrue laugh, softening laugh).  
**Step 7:** Detect the facial expressions during laughter.  
**eyelids**(closed, narrowed, open).  
**cheeks**(raised, not raised).  
**lip concerns**(raised, straightly stretched, lowered).  
**Step 8:** Detect the head and upper body motion during laughter.  
**head**(no motion, up, down, left or right up- down titled nod, others(including motions synchronized with motions like upper-body)).  
**upper body**(no motion, front, back, up, down, left or right, titled, turn, others(including motions synchronized with other motions like head and arms)).  
**Step 9:** End.

Fig. 2. The AHR Generates Laughing Gestures during Dialogue Information.

### B. Adaptive Occurrence

Adaptive occurrence represents adaptiveness as a key characteristic of oral assessment quality. Regarding AHRs in oral assessment, perceived adaptiveness may be expressed as the extent to which students consider the AHR to keep in touch with their individual learning assessment needs. For example, an AHR selects to begin a dialogue with the student to offer assistance, as depicted in Algorithm 3 in Fig. 3. According to Algorithm 3, student assistance suggests more known keywords to help the AHR understand the student-robot dialogue. Then, the AHR solves the problem itself, or student assistance can offer more additional information to enhance the

AHR's understanding. The problem includes understanding keywords or informing unclear messages not inserted into loaded sound files. This assistance provides an appropriate learning pace and introduces personal feedback during the oral assessment.

Furthermore, an AHR can address the unclear content in the oral assessment, as depicted in Algorithm 4 in Fig. 4.

Concerning the social behavior of AHR is a key facet of adaptiveness. The AHR stimulates social interaction in oral assessment with students by calling students by their name and capturing a photo, as depicted in Algorithm 5 in Fig. 5.

**Algorithm 3: AHR starts a dialogue for student's assistance**  
**Step 1:** Start.  
**Step 2:** AHR identifies a problem in student-robot dialogue.  
**Step 3:** AHR searches in loaded sound files on it.  
**Step 4:** If AHR finds the answer  
    **Then** AHR completes the student-robot dialogue.  
    **Else** AHR asks for student assistance.  
  
**Step 5:** Student's assistance provides AHR with the required support.  
**Step 6:** If AHR solves the problem  
    **Then** AHR completes the student-robot dialogue.  
    **Else** AHR requests additional support.  
  
**Step 7:** AHR tries again to solve the problem after student support.  
**Step 8:** AHR solves the problem thanks to Student assistance.  
**Step 9:** End.

Fig. 3. The AHR Starts a Dialogue for Student Assistance.

**Algorithm 4: AHR adapts the dialogue content.**  
**Step 1:** Start.  
**Step 2:** AHR greets students.  
**Step 3:** AHR introduces itself and the educational institution.  
**Step 4:** AHR explains its role in the dialogue which includes to clarify the aim of the dialogue.  
**Step 5:** AHR receives questions from students and provides them with answers.  
**Step 6:** If AHR detects any misunderstanding in the students' dialogue.  
    **Then** AHR alerts their attention and provides them with complete right explanations.  
  
**Step 7:** AHR supports its dialogue with suitable references from the university's library or internet resources or practical examples.  
**Step 8:** AHR updates its dialogue regularly to develop its skills in student-robot interaction.  
**Step 9:** AHR supports student dialogue by using gestures and expressing empathy motions.  
**Step 10:** End.

Fig. 4. The AHR Adapts the Dialogue Content.

**Algorithm 5: AHR adapts social behavior with students.**  
**Step 1:** Start.  
**Step 2:** AHR asks the student about his/her name.  
**Step 3:** AHR greets the student.  
**Step 4:** AHR captures a photo of the student.  
**Step 5:** AHR determines the student's face and performs face recognition.  
**Step 6:** AHR extracts the face features.  
**Step 7:** AHR predicts the student's age and emotional state.  
    (e.g. happy, unhappy, worried, sad, anxious).  
**Step 8:** If there is an academic database about the students' degrees or personal profiles.  
    **Then** AHR can be reached and collects the whole information about the student.  
  
**Step 9:** AHR confirms the learning processes(e.g. oral assessment process,  
    suggests some activities according to the social behavior of students or emotional state).  
**Step 10:** End.

Fig. 5. The AHR Adapts Social behavior with Students.

### C. Persuasion

According to Fig. 6, AHR provides students with more trust and persuasion through its physical shape like humans, which adds more flexibility and comfort during student-robot interaction. Also, the AHR design provides feedback based on social norms and friendship relations. Moreover, the AHR introduces a personalized service with the mutual gaze communicated along with contextual information.

Additionally, it incorporated traditional communicative processes, such as jokes. Through jokes, the AHR can deal with students with double-friendly interaction. Consequently, jokes can motivate students toward effective cooperation and create a community-centered learning environment. Also, jokes encourage students in the learning process and enrich social interaction.



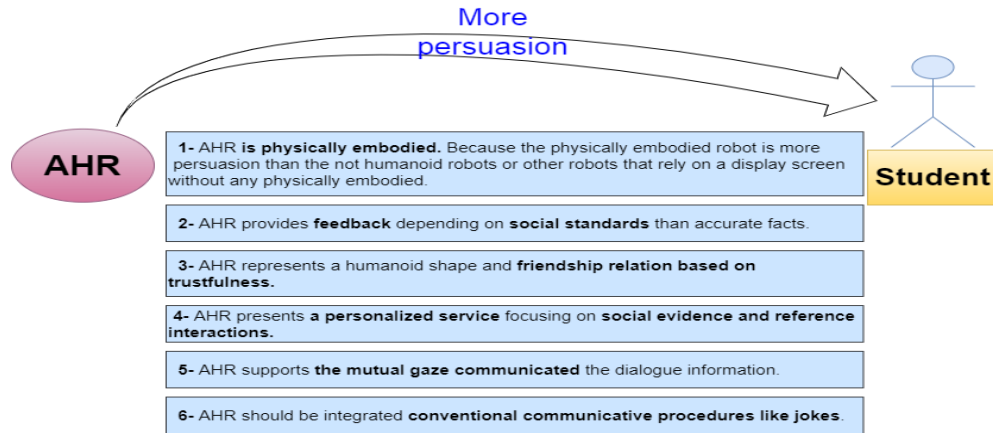


Fig. 6. Persuasion Features of the AHR.

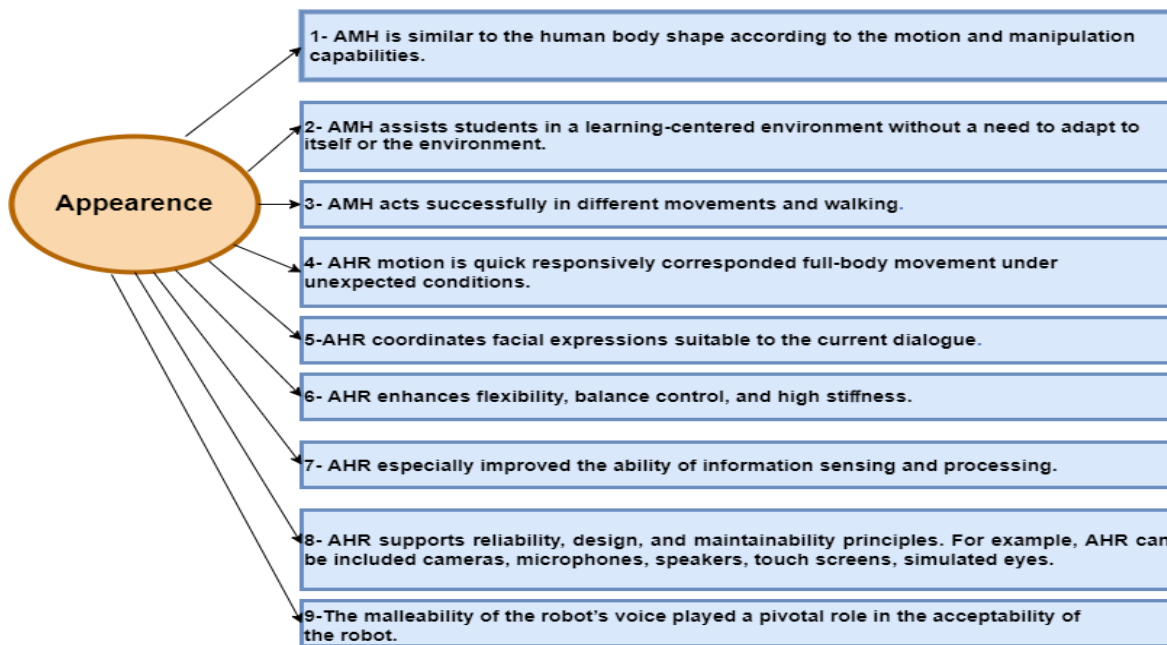


Fig. 7. External Appearance of the AHR.

#### D. External Appearance

AHR appearance should be designed to correspond to dialogue requirements in oral assessments. This appearance can be described as flexible. AHR dialogue contains different alternatives for each social interaction of dialogue, such as a greeting, end dialogue statement, jokes, and different keywords of various intentions. Fig. 7 explains the appearance of the AHR in detail.

### VII. DISCUSSION

To answer the research questions in this study according to the above analysis, we think it is possible for the robot to act as a tutor and automatically guide freshmen to conduct a group test.

We expect robots to judge students' answers fairly. In the past, when human teachers conducted oral tests, they sometimes let students pass the assessment with a looser standard. In addition, human teachers cannot maintain a certain concentration, objectivity and fairly for a long time, and robots can easily do it. However, we can find in the robot was not sensitive to students answers. In fact, sometimes the students should to try several times before the robot could receive the sound and recognize it. Although some monosyllabic words were heard by the robot, there may be some problems because of the pronunciation problems.

For example, since the pronunciation of numbers is very basic, this can be a blow to students. In the first case, students' pronunciation will affect the correct answer rate on the test. From the perspective of language teachers Technical problems can be overcome. Where, we suggest that when using a robot

to conduct the oral test, the robot need to practice pronunciation before conducting formal test. It also could give the students the opportunity to pronounce before the test. Besides, the teacher could modify the way that students answer questions (e.g., by answering with whole sentences instead of just numbers) to make up for the robot's lack of recognition of monosyllables. A further test can be conducted by native speakers to compare the reception of sound to find out if this situation is mainly due to the pronunciation of foreign language learners or the experimental environment. Also about the design of the test, the robot can say the correct answer after each question to let the students know the correct answer or pronunciation immediately.

On the other side, Human-robot interaction also can be enhanced peer-peer interaction. When someone answers right, the peers cheer. Through their peers' affirmation, students can improve self-confidence or reduce the sense of distrust of their own pronunciation. In addition, in group learning students can learn important lessons from the answers made by other students.

### VIII. CONCLUSION AND FUTURE WORK

Humanoid robots play a catalyst role in higher education because they include human-like forms that support student satisfaction. They also encourage students to generate contemporary fluencies. The prosperous performance of emerging AHRs in learning needs cognizance of the investment principles on the part of those carrying out the performance. Oral assessment is a critical process in higher education. Consequently, oral assessment supports active learning through its empowerment of creating confidence among teaching staff and students during the oral assessment process of AHR technology. That is, in turn, it may bring professors and higher education more confidence about the integration of AHRs technology into their strategy. This paper has discussed an algorithm for applying AHR in oral assessments in higher education. This algorithm also depends on four perceived characteristics: adaptive occurrence, friendly existence, persuasion, and external appearance. Moreover, it presents the experiences of human-robot interaction.

In future work, we intend to discover substantial distinctions between oral assessment guided by the teaching staff and AHRs. Also, we hope to develop the proposed algorithm by adding more characteristics. We desire to be able to use the AHR to direct oral tests so that professors can exchange roles and monitor and assess students' complete ability and change the way the oral test is conducted.

#### REFERENCES

- [1] Pandey AK, Gelin R. Humanoid robots in education: a short review. *Humanoid Robotics: a Reference*. 2017;1-16.
- [2] Holler J, Shovelton H, Beattie G. Do iconic hand gestures really contribute to the communication of semantic information in a face-to-face context? *Journal of Nonverbal Behavior*. 2019;33(2):73-88.
- [3] Salem M, Eyssel F, Rohlfsing K, Kopp S, Joubin F. To err is human (-like): effects of robot gesture on perceived anthropomorphism and likability. *International Journal of Social Robotics*. 2013;5(3):313-323.
- [4] Yoon Y, Ko WR, Jang M, Lee J, Kim J, Lee G. Robots learn social skills: end-to-end learning of co-speech gesture generation for humanoid robots. In *2019 International Conference on Robotics and Automation (ICRA)*. IEEE. 2019 May;4303-4309.
- [5] Lin P, Abney K, Bekey GA, editors. *Robot ethics: the ethical and social implications of robotics*. MIT Press. c2014.
- [6] Tuna G, Tuna A, Ahmetoglu E, Kuscu H. (A survey on the use of humanoid robots in primary education: prospects, research challenges and future research directions. *Cypriot Journal of Educational Sciences*. 2019;14(3):361-373.
- [7] Randall N. A survey of robot-assisted language learning (RALL). *ACM Transactions on Human-Robot Interaction (THRI)*. 2019;9(1):1-36.
- [8] Li H, Yang D, Shiota Y. Exploring the possibility of using a humanoid robot as a tutor and oral test Proctor in Chinese as a foreign language. In: *Expanding global horizons through technology enhanced language learning*, Springer, Singapore; c.2021. p. 113-129.
- [9] Van den Bergh R, Verhagen J, Oudgenoeg-Paz O, Van der Ven S, Leseman P. Social robots for language learning: a review. *Review of Educational Research*. 2019;89(2):259-295.
- [10] Mende M, Scott ML, van Doorn J, Grewal D, Shanks I. Service robots rising: how humanoid robots influence service experiences and elicit compensatory consumer responses. *Journal of Marketing Research*. 2019;56(4):535-556.
- [11] Stock-Homburg R, Hannig M, Lilienthal L. Conversational flow in human-bot interactions at the workplace: comparing humanoid and Android robots. In: *International conference on social robotics*. Springer, Cham; c2020 Nov. (p. 578-589).
- [12] Rojas-Quintero JA, Rodríguez-Liñán MC. A literature review of sensor heads for humanoid robots. *Robotics and Autonomous Systems*. 2021;143:103834.
- [13] Pajaziti A, Bajrami X, Pula G. *Communication and Interaction between Humanoid Robots and Humans*. c2021.
- [14] Filippini C, Perpetuini D, Cardone D, Chiarelli AM, Merla A. Thermal infrared imaging-based affective computing and its application to facilitate human robot interaction: A review. *Applied Sciences*. 2020;10(8):292
- [15] Aburlasos VG, Dardani C, Dimitrova M, Amanatiadis A. (2018, January). Multi-robot engagement in special education: a preliminary study in autism. In: *2018 IEEE International Conference on Consumer Electronics (ICCE) IEEE*. p. 1-2.
- [16] Mazhar O, Navarro B, Ramdani S, Passama R, Cherubini A. A real-time human-robot interaction framework with robust background invariant hand gesture detection. *Robotics and Computer-Integrated Manufacturing*. 2019;60:34-48.
- [17] Alemi M, Meghdari A, Haeri NS. Young EFL learners' attitude towards RALL: An observational study focusing on motivation, anxiety, and interaction. In *International Conference on Social Robotics*. Springer, Cham; c2017 Nov. p. 252-261.
- [18] Henschel A, Hortensius R, Cross ES. Social cognition in the age of human-robot interaction. *Trends in Neurosciences*. 2020;43(6):373-384.
- [19] Hussein A, Gaber MM, Elyan E, Jayne C. Imitation learning: a survey of learning methods. *ACM Computing Surveys (CSUR)*. 2017;50(2):1-35.
- [20] Hua J, Zeng L, Li G, Ju Z. Learning for a robot: deep reinforcement learning, imitation learning, transfer learning. *Sensors*. 2021;21(4):1278.
- [21] Chen X, He B, Katz GE. Towards human-like learning dynamics in a simulated humanoid robot for improved human-machine teaming. In: *International Conference on Human-Computer Interaction*. Springer, Cham.; 2022. p. 225-241.
- [22] Memon MA, Joughin GR, Memon B. Oral assessment and postgraduate medical examinations: establishing conditions for validity, reliability and fairness. *Advances in Health Sciences Education*. 2010;15(2):277-289.
- [23] Memar AH, Esfahani ET. Objective assessment of human workload in physical human-robot cooperation using brain monitoring. *ACM Transactions on Human-Robot Interaction (THRI)*. 2019;9(2):1-21.
- [24] Guggemos J, Seufert S, Sonderegger S. Humanoid robots in higher education: evaluating the acceptance of Pepper in the context of an academic writing course using the UTAUT. *British Journal of Educational Technology*. 2020;51(5):1864-1883.
- [25] Fang B, Jia S, Guo D, Xu M, Wen S, Sun F. Survey of imitation learning for robotic manipulation. *International Journal of Intelligent Robotics and Applications*. 2019;3(4):362-369.

# Classifiers Combination for Efficient Masked Face Recognition

Kebir Marwa, Ouni Kais

Research Laboratory Smart Electricity & ICT, SE&ICT Lab., LR18ES44

National School of Engineers of Carthage

University of Carthage, Tunisia

**Abstract**—This study was developed following the upheaval caused by the spread of the Coronavirus around the world. This global crisis greatly affects security systems based on facial recognition given the obligation to wear a mask. This latter, camouflages the entire lower part of the face, which is therefore a great source of information for the recognition operation. In this article, we have implemented three different pre-trained feature extractor models. These models have been improved by implementing the well-known Support Vector Machines (SVM) to reinforce the classification task. Among the investigated architectures, the FaceNet feature extraction model shows remarkable results on both databases with a recognition rate equal to 90% on RMFD and a little lower on SMFD with 88.57%. Following these simulations, we have proposed a combination of classifiers (SVM-KNN) that would prove a remarkable improvement and a significant increase in the accuracy rate of the selected model with almost 4%.

**Keywords**—Masked faces; deep learning; AlexNet; ResNet50; FaceNet; classifiers combination

## I. INTRODUCTION

According to the World Health Organization (WHO) [1], there have been 517.648.631 confirmed cases of COVID-19, including 6.261.708 deaths, until May 13th, 2022. That is considered an astonishing number after three years of the virus's appearance and despite all precautions taken. Therefore, the Centers for Disease Control and Prevention (CDC) [2] emphasize social distancing and obligation to wear masks in order to minimize contamination and reduce the hazardousness of this virus. Except that, wearing a mask causes the performance degradation of security systems based on the identification of people by their faces, since the mask hides a large part of the face, hence the loss of a large amount of information. Thus, existent techniques for faces recognition implemented before this crisis, must have an improvement and some adjustment to live up to the expectations of users of face-based security systems, as stated by the National Institute Of Standards and Technology (NIST) [3]. The performance of facial recognition algorithms submitted before March 2020, when the World Health Organization declared a global pandemic, was examined in a previous NIST report published in July. The error rate of these pre-pandemic algorithms was found to be between 5% and 50% in this first investigation, which confirms that these systems have become ineffective.

The degradation of identification systems, causes several problems. So, frauds and wanted persons take advantage of the mask, for illegal immigration and crimes commitment without being recognized. Community access control and face

identification have become almost an impossible mission when a grand portion of the face is covered up by a mask. Due to these issues, face masks have essentially challenged existing face recognition strategies.

The epidemic situation, ensures the emergence of two new axes of research: [4]

- Face mask detection: consists of checking whether the individual is wearing a mask or not, and it is an interesting task in public squares and areas with fulls, where wearing a mask is mandatory.
- Masked face recognition (MFR): is used to identify people wearing masks on the basis of the remaining part of the face (the eyes and the forehead parts)

Our interest in this paper is the second axis. we have implemented a recent technique to identify masked faces using a deep learning-based method for extracting features. To train our model, two databases are used: Simulated Masked Faces Recognition database(SMFRD) and Real-World Masked Faces Database(RWMFD) presented in [5], specially designed to evaluate the performance of masked faces recognition methods.

An interesting preliminary phase in the recognition operation is the pre-processing of the images. However, the quality of the detection influences the accuracy of the identification, so we chose the MTCNN algorithm to have an exact and correct detection. Regarding the feature extraction task, we have opted for three pretrained models that are very recognized in the field of facial recognition, which are AlexNet, ResNet50 and FaceNet. For the classification process we have chosen a classifier which has proven a huge success in image classification, it is the Support Vector Machine(SVM). Finally, we have proposed, a classifiers combination (SVM-KNN) to enhance the performance of the classification process.

This study is organized as follows: The Section 2 provides the related works about masked faces recognition. While Section 3 highlights the motivation and the contribution of the paper. Section 4 discusses the state-of-the-art methods. Section 5 presents the used method. The remainder of the paper stated the experiments and the concluding statements based on experimental results.

## II. RELATED WORKS

Obviously, the issue of partially hidden faces has existed for a long time, since there are several factors other than wearing a mask such as a beard or mustache, sunglasses or

even makeup that changes the original features for masquerade reasons. But recently, the obligation to wear a mask makes the problem even worse and the dilemma of recognizing occluded faces has become at the head of research in computer vision. Consequently, significant increase in MFR research effort, extending existing MFR methods and yielding promising accuracy results. Then, search across major digital libraries to track growing research interest in the mission of occluded face recognition (OFR). A series of search strings are formulated to find leading repositories the article covers the use of deep learning techniques only in the context of face-based recognition. MFR article search results retrieved from Web of Science, Scopus, IEEE Xplore, Wiley, Ei Compendex, ACM Digital Libraries and EBSCOhost. These warehouses include articles on recent popular seminars, journals and conferences articles from five years.

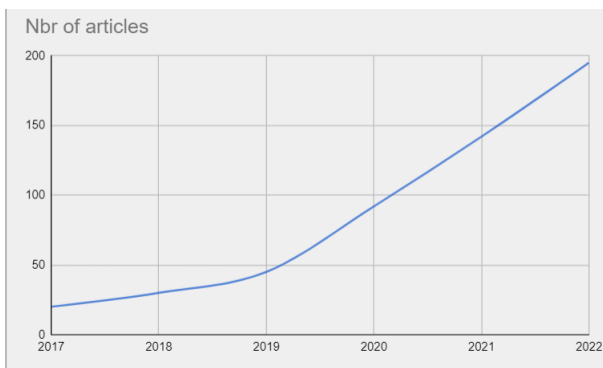


Fig. 1. Diagram of Research Efforts on MFR from 2017 to 2022.

As shown by Fig. 1, the importance of research in the field of Occluded and Masked Faces recognition (OFR and MFR) witnesses a noticeable increase in parallel to exhausting researches [6] [7] [8].

Except that, this diversity does not mean the effectiveness of the applied techniques, because until now there is no masked face recognition method whose performance exceeds or equals that of unmasked face identification techniques. As a result, the achievement of the methods used post-pandemic remain unsatisfactory for real-time systems and high-traffic sites with high security requirements. As of late, researches have proposed many useful methods. They basically consist of three categories:[9]

- Generate a typical model of the occlusion issue(restoration model),
- Occlusion removal approach,
- Deep learning-based approaches.

#### A. Restoration Model

This approach consists of generating the hidden part of the face considering that the nose, mouth and chin carry a large amount of information. For the improvement of the recognition efficiency reasons and to generate the lost features of face, there are two restoration models: robust structured error coding and robust subspace regression [9].

- Robust structured error: The occlusion produced by the mask presents a spatial continuity. By this, an error caused has a specific spatial structure. This makes the reconstruction of the low-rank structure of the face image from the data damaged by occlusion necessary, to have a correct identification and minimize the rate of false positives and true negatives. For instances, authors in the literature [10] presented an improved robust principal component analysis (RPCA) method. At the beginning, the method consists in decomposing the learning matrix  $M$  by a lower rank matrix, which ensures obtaining a lower rank content matrix  $L$  and a sparse content matrix  $S$ . In this way, the recovery of the lower subspaces of the training sample is achieved. This method considers how to restore low-ranked structures from training samples that are error-prone but sparsely structured. This effectively suppresses the effects of sparse noise and provides strong efficiency. In order to increase interclass information between low-ranked matrices of different face categories, reference [9] has expressed all training patterns as observation matrix  $D$ . After decomposing the matrix  $D$ , we get a low-ranked matrix  $A$  with no occlusion and a sparse error matrix  $E$ . The RPCA is applied to the submatrix  $A$  and the resulting subspace is used as an occlusion dictionary for facial images. The image reconstruction was then identified and the error size was classified according to the sparse representation classification and occlusion dictionary.[9]

- Robust subspace regression: This model is generated by projecting high-dimensional feature data from different categories of facial images onto a low-dimensional subspace. Next, an independent subspace is set in the occlusion part, and the occlusion of the face image is expressed using the existing dictionary atom to realize a powerful recognition effect of the occlusion face. Currently, robust subspace solutions for occluded face detection primarily include sparse representation, collaborative representation, and obstruction dictionary learning.

#### B. Occlusion Removal Approach

This model aims to estimate the position of occlusion through two error indices. The first is in the form of local similarity error between the original image and the partially occluded image, while the second aims to the spatial local error caused by the occlusion. Otherwise, the method consists in locating the hidden areas of the face and eliminating them completely from the feature extraction and classification process. In this context, one of the most famous approaches is that based on segmentation. According to the literature [11], authors segmented the face in small local zones. These latter, contain the occluded part to be eliminated and which will be detected by the support vector machine (SVM). Then the last phase is to use a mean-based weight matrix for face identification.

Several techniques have been proposed to remove the concealed part of the face, including the exemplar based Image in-painting technique proposed in [12]. As well as, the

structural similarity index measure and principal component analysis technique [13].

### C. Deep Learning-Based Approaches

Currently, Deep learning has proven huge success in several areas, especially in FR. This comes down to the efficiency of deep features compared to others that are over shallow. The most recognized research works in occlusion face are based on this type of features, taking the example in [14], an efficient partial face recognition approach was proposed, this is Dynamic Feature Matching (DFM) approach, based on the combination of the Fully Convolutional Network (FCN) with Sparse Representation Classification (SRC). The method is intended to recognize partial faces of arbitrary size. Another model proposed to identify hidden faces called BoostGAN model [15]. The main idea of this model is to use the occluded face to elaborate the non-occluded face and this latter will then be used to recognize the person. Except that, in the case of large occluding surfaces such as face masks, GAN-based methods are hard to regenerate the details of the key points on the visage.

Table I presents the performance of some MFR methods.

TABLE I. SUMMARY OF SOME MFR METHODS PERFORMANCE

References	Model	Dataset	Accuracy
GuiLing, W.[9]	Mask separation	RMFRD	95.22
Priya, G.[11]	MBWM-SVM	GTAV	94.75
He,L.[14]	Dynamic Feature Matching	CASIA-NIR-Distance	94.96

### III. MOTIVATION AND CONTRIBUTION OF THE PAPER

The failure of current methods to correctly identify masked faces in the same way as non-masked face recognition techniques motivates us to explore a new solution to overcome this shortcoming. Drawing on the significant performance and strong light resistance of CNN-based methods, facial expression variations and face occlusion. In this study, we have proposed an occlusion removal approach with transfer learning model to solve the masked face problem noticing during the COVID-19 pandemic.

To the best of our knowledge, the restoration model approach isn't truly a great selection. This procedure endures from a few issues, particularly the difficulty of implementing it and culminating the comes about. As well as, it could be a prepare that devours an expansive execution time. With that, several researchers use this method as is the case in the literature [16], contrary of our choice. The disadvantages of this approach, inspired us to base our considerations on the occlusion removal approach since it is a compelling and simple strategy. As well as, pre-trained model for feature extraction.

Other side, face detection has a huge influence on the quality of recognition. As of late, a few face tracking strategies have shown up and have demonstrated tall execution, such as the viola and Jones detector, which has gotten to be a reference in object detection and particularly for face. A bit more recently, it appears the famous MTCNN detector [17] which has proven great efficiency and speed of execution. But that the major progression of machine learning recently, gives us with unused choices, like the modern finder which

made a boom in computer vision is the mediapipe algorithm [18]. MediaPipe face detector is an ultra-fast solution with six landmarks and multi-face support. It is an excellent detector for streaming videos, but it is not yet used for static images. What motives us to use the MTCNN detector.

So the contribution of this paper lies in three points:

The first consists in adapting models usually used for the recognition of unmasked faces for masked faces. This adaptation was made by training the three models AlexNet, ResNet50 and FaceNet by databases specially designed for masked faces and adjusting the parameters of each model in order to improve their performances.

The second contribution aims to improve classification task with the inserting of SVM classifier for each model.

While the third contribution consists in a classifiers combination (SVM-KNN) that considerably improves the obtained results.

### IV. STATE OF-THE-ART METHODS

The ordinary and common structure for all face recognition methods is composed of a face detection, a feature extraction and a classifier. Through this paper, we have presented the races that we followed, starting with face cropping.

#### A. Face Detector: MTCNN

Currently, MTCNN or Multi-task convolutional neural networks, is the most popular and rigorous face detection solution. It is composed of three cascaded neural networks known by P-Net, R-Net and O-Net [17].

- **P-Net** stands for **Proposed network**: It searches for faces in frames of size  $12 \times 12$ . The task of this network is to achieve rapid results.
- **R-Net** presents **Refined network**: Its structure is deeper than Pnet. Despite all candidates originating from the previous network will be fed to R-Net, a large number of candidates are already eliminated by the first network P-Net.
- **O-Net** comes from **Output network**: briefly returns the bounding box (face area) and face landmark locations.

These three cascaded neural networks are presented in Fig. 2.

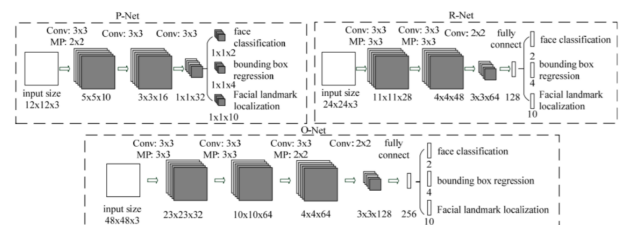


Fig. 2. Architecture of the Three Cascaded Neural Networks of MTCNN [17]

The MTCNN model detects five landmarks on the face, which are the left eye, right eye, nose, and two corners of the mouth. This model proves high face detection accuracy.



### B. Masked Faces Databases

In this section, we have presented several benchmark datasets used in the literature to evaluate MFR techniques [19].

Starting with the most popular and used databases in this field which are the RWMFRD (Real-world masked faces recognition dataset), the Masked Face Detection Dataset (MFDD) and Simulated Masked Face Recognition Dataset (SMFRD) introduced in the same article [5].

As regards **MFDD**, it contains 24,771 masked face images, which allows the implemented model to accurately detect faces hidden by masks.

As for **RMFRD**, the largest existent database for MFR, since it contains 5,000 images of 525 people wearing masks, and 90,000 images of the same people without masks. This database was used in [20], in the context of face images that were not acceptable due to an incorrect match were manually removed. In addition, the right face areas have been cropped using semi-automatic annotation techniques, such as LabelImg and LabelMe, Fig.3 displays sample images from RMFRD.



Fig. 3. Sample Images from RMFRD.

For diversification reasons, **SMFRD** has been developed, it contains 500,000 images of synthetically masked faces of 10,000 people collected from the Internet and Fig. 4 illustrate some images from this dataset.

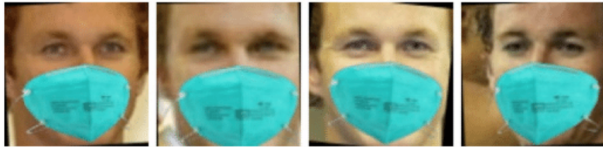


Fig. 4. Sample Images from SMFRD.

The Synthetic face-occluded dataset (**SFOD**) [21] was elaborated using published data records from CelebA and CelebA-HQ [22]. CelebA-HQ is a large-scale facial attribute dataset containing over 30,000 celebrities. Each face image is cropped and roughly aligned based on the eye position. The occlusions were aggregated by five common non-facial objects: hands, masks, sunglasses, glasses, and microphones. Over 40 different types of objects were used in different sizes, shapes, colors and textures. In addition, non-face objects were randomly placed on the face.

The Masked Face Segmentation and Recognition dataset (**MFSRD**) [23] is composed of two parts. The first part consists of 9742 images of masked faces that have been collected from Internet with hand-labeled masked segmentation annotation. The second part contains 11,615 images of 1004 identities, of which 704 are collected from the real world and the rest of the images are collected from the Internet, where each identity has at least images with and without masks. Celebrities in Frontal Profile in the Wild (**CFP**) [24] includes the faces of 500

celebrities in face and profile view. Two verification protocols with 7000 comparisons, each presented: one compares only frontal faces (FF) and the other compares (FF) and profile faces (FP).

Both, Masked Face Verification (**MFV**) and Masked Face identification (**MFI**) are presented in [25], the first one contains 400 pairs for 200 identities while the second includes 4916 images of 669 identities.

The **LFW-SM** [26] variant database contains a simulated mask that extends the LFW dataset and contains 13,233 images from 5749 individuals. Through Fig. 5, we have presented sample images from LFW-SM.



Fig. 5. Sample Images from LFW-SM.

Several MFR techniques used the VGGFace2 [27] dataset for training, which consists of 3 million images of 9131 people with over 362 images per person. From this database derives the Masked faces dataset **VGGFace2-m** [28], it contains over 3.3 million images of 9131 identities. Table II shows the main characteristics of the dataset used in the masked face recognition task.

TABLE II. SUMMARY OF THE MFR BENCHMARKING DATASETS

Database	Size	Identities	Type of masks
RMFRD	95,000	525	Real-world
SMFRD	500,000	10,000	Synthetic
MFSRD	11,615	1004	Real-world/synthetic
MFV	400	200	Synthetic
MFI	4916	669	Synthetic
LFW-SM	13,233	5749	Synthetic
VGG-Face2-m	3.3M	9131	Synthetic

### C. Feature Extraction

One of the most crucial steps in the facial recognition process is feature extraction. It consists of extricating a set of features that are discriminative enough to represent and learn key facial attributes such as eyes, mouth, nose, and texture. In the presence of partial occlusion, especially that produced by the face mask, this process becomes more complex and current facial recognition systems need to be adjusted to extract representative and robust facial features. There are two categories of feature extraction, the first is a **shallow feature extraction** which is a classical technique explicitly forming a set of features fabricated with low optimization or learning mechanisms. The most popular methods of this category are Histogram of Oriented gradient (HOG), LBPs and codebooks [29]. In recognition tasks for unmasking face, these algorithms achieve considerable accuracy and remarkable robustness against a variety of facial changes such as lighting, rotation, scaling and translation. But this is not the case for masked faces, great degradation was observed.



The second category is the **deep feature extraction**. One of the most efficient neural networks in the field of face recognition is the Convolutional Neural Network (CNN), it has shown preponderance in a wide range of applications, such as image classification, retrieval and detection of objects. CNN-based models have been widely deployed and trained on many large-scale face datasets [30] [31] [32].

Several pre-trained architectures are recognized in the field of FR and have proven remarkable success, especially for feature extraction. To choose the most suitable extractor for our application, we have made an overview of the most cited models in the literature. AlexNet [33] is a famous model that ensures to reduce the training time and minimize the errors even on large datasets [34]. Fig. 6 shows the architecture of AlexNet.

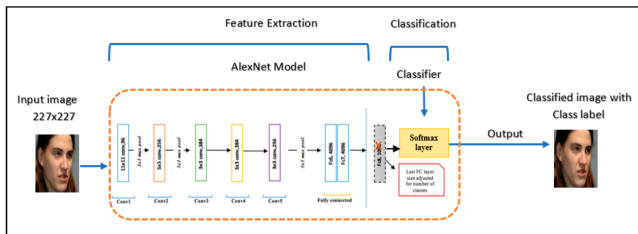


Fig. 6. Architecture of AlexNet [35].

Two other popular CNN-based models were presented in [36], VGG16 and VGG19 have been used in various computer vision applications, especially facial recognition. Despite achieving considerable accuracy, they endure from training time and complexity [37].

For the most complex identification missions, as in the case of masked faces, it is preferable to be processed by deeper neural networks like residual network (ResNet) [38]. This model achieves outstanding performance and accuracy, due to the stack of additional layers. These extra layers must be determined empirically to control for any deterioration in the performance of the model. The architecture of ResNet50 is presented in Fig. 7.

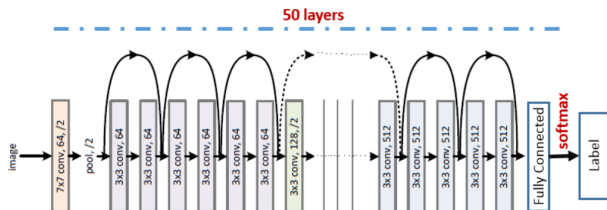


Fig. 7. Architecture of ResNet50 [39].

MobileNet [40] also is considered as one of the earliest deep neural networks, which mainly depends on a simple architecture. Its architecture exhibits high performance with hyperparameters and fast model calculations [41].

We cannot pass without mentioning Inception [42] and its variations [43], their innovation is that they use modules or blocks to build networks that contain folding layers instead of stacking them. Xception [44] is an extreme Inception

version that replaces Inception modules with deeply separable convolutions.

FaceNet [45], presented by Google researches, it is a famous pre-trained model that has proven very remarkable results. Fig. 8 shows the architecture of FaceNet.



Fig. 8. Architecture of FaceNet [45].

Through Table III we have presented a summary of the pre-trained CNN-based model.

TABLE III. SUMMARY OF THE PRE-TRAINED CNN BASED MODEL

Model	Variants	Trainable param	Conv layers	Total layers
AlexNet	-	62 M	5	8
VGG	VGG16	138 M	13	16
	VGG19	143 M	16	19
ResNet	ResNet50	25 M	48	50
	ResNet101	44 M	99	101
MobileNet	MobileNet	13 M	28	30
	MobileNet-v2	3.5 M	-	53
Inception	GoogleNet	7 M	22	27
	IncepV2	56 M	22	48
	IncepV3	24 M	22	48
	IncepV4	43 M	-	164
	Incep-ResNet-V2	56 M	-	164
Xception	-	23 M	36	71
FaceNet	-	140M	22	27

#### D. Classification

Many classifiers have been mentioned through literature, given the importance of classification in improving the performance of facial recognition systems.

As far as we are aware, the two most popular classifiers in facial recognition are Support Vector Machine (SVM) and K-Nearest Neighbour (K-NN) [46]

1) SVM: Support vector machine (SVM) is a supervised machine learning that can be used for classifications or regression problems. For the issues of multiclass classification it exists two distinguished approaches, the first is one-against-one and the second is one-against-all approach. Kernel functions are used for separation between classes for higher dimensional feature spaces. These Kernel functions are able to transform a non-linear distinguishable problem into a linear distinguishable one and projecting data into the feature space which ensure to find the optimal separating hyper plane [51].

2) KNN: K-Nearest Neighbor, a popular technique for classifying objects based on nearest training samples in feature space. This training samples are vectors with a class label for each one. The principle of the technique aims to compute the distance of the test sample to every training sample and keeping the k closest training samples (Where k designate positive integer). Then several distance functions used in the KNN algorithm, but the best methods are Euclidean distance [52].

3) *SVM-KNN*: Overall, the combination of classifiers is a relatively new technique. It can be considered as an optimization problem for minimizing classification errors and takes as input the outputs of M classifiers and generates the final N classes [53]. Classifier combination is more efficient, especially when the classifiers are different. There are two types of combination:

Features association using similar classifiers and decision association resulted from dissimilar classifiers [54]. The second type of combination is our choice since SVM and k-NN are two dissimilar classifiers.

On the other hand, classifiers can provide three types of outputs: The measure, class and rang types [55]. Depending on these types of output the combination may be:

- In the abstract stage built on voting methods.
- In the rank stage when the outputs are labels classified by a reducing weight.
- In the measurement level if the outputs are labels combined with confidence values.

For our work, we have used the majority vote in combination.

### E. Evaluation Metrics

In order to assess the cogency of implemented models for masked face recognition, we have opted for some evaluation parameters, such precision, recall, F1-score and accuracy metrics. These evaluation metrics were calculated as follow [49]:

$$Accuracy = \frac{T_N + T_P}{T_S} \quad (1)$$

$$Precision = \frac{T_P}{T_P + F_P} \quad (2)$$

$$Recall = \frac{T_P}{T_P + F_N} \quad (3)$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} + score \quad (4)$$

Where  $T_P$ ,  $T_S$ ,  $F_P$ ,  $T_N$  and  $F_N$  designate respectively True positive, total samples, false positive, true negative, and false negative. For our work we have used Accuracy, precision and recall rate.

## V. USED METHOD

Our method consists of three primary scenarios (presented in Fig. 9), where we used three different pre-trained CNN-based models like feature extractors and the SVM is cascaded for classification. The model that has proven the best results will be improved by the combination of a second classifier which will be the K-NN.

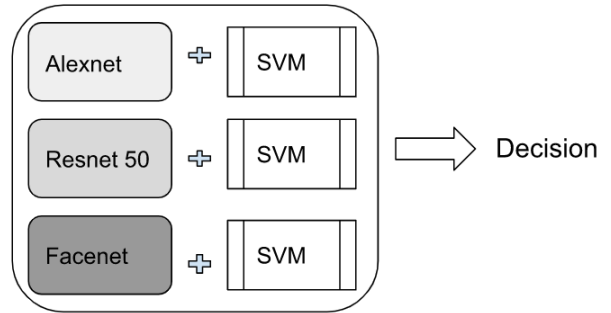


Fig. 9. Architecture of our Primary Scenarios.

### A. AlexNet and SVM

We have used the well-known SMFD and RMFD databases specially designed to evaluate masked facial recognition systems. Data from each base were divided into 80% for training and 20% for testing and random distribution to avert biased results. Then, the images used are a mix of masked and unmasked faces which ensures good feedback.

Before we start, we need to preprocess the input images to fit the AlexNet model. Input image should be with a dimension equal to  $227 \times 277 \times 3$  pixels. For SMFD database, the image contains only the parts of the face we are interested in, so we don't need face detector, we just need to convert the size of the images from  $128 \times 128 \times 3$  to  $227 \times 227 \times 3$ . Unlike the RMFD base, we use MTCNN to frame the faces. It is important to note that the model is made up of many layers, but not all of these layers are essential for feature extraction. For instance, the first layer deals with the extraction of features such as edges and points. As it is presented in Fig. 10.

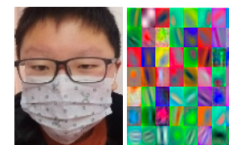


Fig. 10. Extracted Features by the First Layer of AlexNet from Original Image on RWMFD.

The table in Fig. 11 shows the different layers of the Alexnet model.

Layer	Number of Kernels	Kernel Size	Stride	Padding	Output Size
Input					$[227 \times 227 \times 3]$
Conv1	96	$11 \times 11 \times 3$	4	-	$[55 \times 55 \times 96]$
Max pool1		$3 \times 3$	2	-	$[27 \times 27 \times 96]$
Norm1					$[27 \times 27 \times 96]$
Conv2	256	$5 \times 5 \times 48$	1	2	$[27 \times 27 \times 256]$
Maxpool2		$3 \times 3$	2	-	$[13 \times 13 \times 256]$
Norm 2					$[13 \times 13 \times 256]$
Conv3	384	$3 \times 3 \times 256$	1	1	$[13 \times 13 \times 384]$
Conv4	384	$3 \times 3 \times 192$	1	1	$[13 \times 13 \times 384]$
Conv5	256	$3 \times 3 \times 192$	1	1	$[13 \times 13 \times 256]$
Max pool3		$3 \times 3$	2	-	$[6 \times 6 \times 256]$
fc6 Rel,U Dropout(0.5)	1				4096
fc 7 Rel,U Dropout(0.5)	1				4096
fc8 softmax	1				1000

Fig. 11. Details of AlexNet Layers [47].

After features extraction, we have equipped the SVM to perform the classification task. The output from ‘fc8’ layer was a 4096-dimensional feature vector. For the SVM kernel function, we used a linear kernel function without optimization. Kernel functions are used to take vector data as input and convert it into the optimal format. We were inspired by this network implemented mainly for unmasked faces as displayed by Fig. 12.

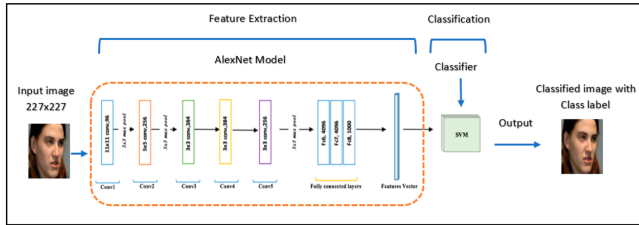


Fig. 12. AlexNet Convolutional Neural Networks with SVM.[35]

**B. ResNet50 and SVM**

Firsthand, Through the implementation operation, we have created an image data store that is useful for data management. Seeing that, the image was read and then loaded into the storage system. Then, the data was split into 80% training and 20% validation using random sampling for averting bias in the outcomes. On the other hand, the ResNet50 network can only process images with  $224 \times 224 \times 3$ , that’s why we resized images to be used in this size. Regarding the next phase, which is feature extraction, there is a final layer named fc1000 found just before the classification layer was used to extract features using the activation method. The activation outputs were aligned in columns to speed up afterwards the SVM training , and the optimizer Adam was used for the training instead of Stochastic Gradient Descent. Fig. 13 presents the ResNet50 model implemented with SVM.

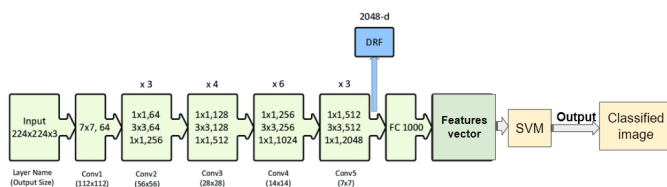


Fig. 13. ResNet50 Convolutional Neural Networks with SVM.

We have adjusted manually different hyper parameter to obtain better outcomes. Starting with the batch size, it is tuned to 32 and we run our algorithm with cross-entropy for 50 epochs.

**C. FaceNet and SVM**

FaceNet model accomplished state-of-the-art results in several benchmark face recognition datasets, specially Labeled Faces in the Wild (LFW) and YouTube Face Database. The model requires as input images sizes  $160 \times 160$  and it contains 22 deep layers and 5 pooling layers, and a global average pooling is used at the end of the last inception module. The Fully connected layer will be used for face description. Elaborated

descriptors become an embedding module for correspondence descriptors. The max operator has been applied to features to develop a one feature vector from a template. The network must be properly tuned to expect a significant boost for the particular task of face recognition and verification. To retrain the FaceNet model, we have to bring a set of masked and unmasked faces[42].

The module includes four branches, the first contains a series of 1x1 local features from the input for learning. While The second branch implements 1 x 1 convolution in order to reduce the input dimensions until 1 x 1 convolution is achieved. This greatly minimises the quantity calculation the network desires. Third branch is coherent with the second branch with 5x5 learning filters. The last branch accomplishes 3x3 max pooling with a Stride of 1x1. Finally, all branches of the Inception module converge and Channel dimensions are linked to each other before being added to the next network, as displayed by Fig. 14 [48].

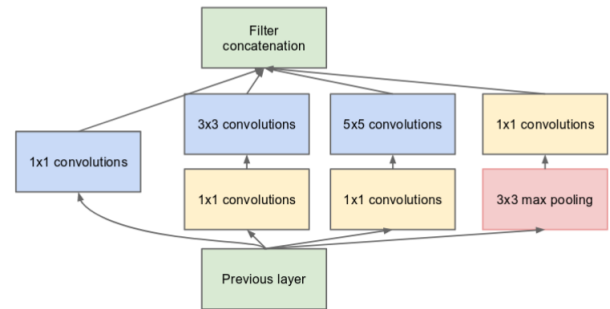


Fig. 14. Illustration of the Inception Module used for FaceNet Pre-Trained Model.

A crucial and next task in the FaceNet model is Face embedding, it consists of the representation of facial feature in the form of a vector. This latter, will be useful for comparison with the other generated vectors for identification of people. Embedding vector will be stored in order to be utilized as an input for the classifier. For this reason, we have to itemize each face in both training and testing database to have the classifier perform embedding and name prediction. It’s necessary that the pixels of the image are normalized to perform the prediction operation. FaceNet architecture contains a batch layer and a deep CNN network. This latter, was supported by the normalization L2 which results face embedding. The face embedding is performed Triplet loss during the training. The triple loss has a minimum distance between an anchor and positive when the identities are the same [49]. Fig. 15 displays the triplet loss training.

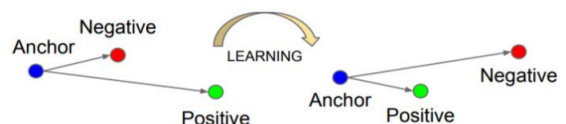


Fig. 15. The Triplet Loss Training[45]

Then, the validation process is integrated to recognize a candidate’s face by performing a classification task within an

integrated support vector machine (SVM). Since its inception, the SVM algorithm has been effectively applied to various classification-related problems. The SVM finds an Hyperplane that performs the classification task of the optimization issues. This maximizes the boundaries between the two classes of a particular input and target pair. The Classification is the result of certain robustness against over-fitting and margins represent class separation efficiency.

## VI. EXPERIMENTS AND RESULTS

### A. Implementation

Through this section, we have presented the experimental results acquired by face recognition using the three deep convolutional neural networks previously mentioned, after they are chained by the famous SVM classifiers, these implementations are based on both MFRD and SMFD databases. Three major experiments in our study were performed to compare performance differences between pre-trained CNN architectures. First, we evaluated the performance when extracting the learned image features from a pre-trained CNN AlexNet, followed by SVM as a classifier. Second, we have realised the same experience with ResNet50. Third, we have evaluated the performance of FaceNet model with SVM. The analysis and evaluation were carried out on the basis of the performance recognition accuracy, precision and recall rate.

Before beginning the training process for the convolutional neural network architectures, a previous pre-processing is required. For all datasets, a rescale is applied to resize the images to a  $227 \times 227$  as input for AlexNet,  $224 \times 224$  as input for ResNet50 and  $160 \times 160$  for FaceNet model.

All experiments were conducted using the platform of Windows with the configuration of AMD Ryzen5-GPU with 16 GB of NVIDIA GEFORCE RTX 3050 TI. Python tool was used to evaluate the method and perform the feature selection and classification task.

Table IV and Table V display the results obtained from the pre-trained models using two different databases.

TABLE IV. EVALUATION OF IMPLEMENTED MODELS USING RMFD

Model	Accuracy%	Precision%	Recall%
AlexNet	88.89	90.00	90.00
ResNet50	84.20	83.650	85.310
FaceNet	90	90	91.5

TABLE V. EVALUATION OF IMPLEMENTED MODELS USING SMFD

Model	Accuracy%	Precision%	Recall %
AlexNet	85.210	88.12	88.12
ResNet50	83.304	81.870	84.10
FaceNet	88.57	88.5	88.6

Since FaceNet has demonstrated considerable robustness with masked faces, we have opted this model to improve this performance by combining the used SVM classifier with the well-known K-NN classifier. The results obtained are presented in Table VI.

TABLE VI. FACENET MODEL WITH CLASSIFIERS COMBINATION

Model	RMFRD	SMFRD
FaceNet with SVM	90%	88.57%
FaceNet with SVM-KNN	94.46%	91.87%

### B. Discussion

As a first observation, the two databases are not simulated in the same way, RMFD provides more considerable results than the SMFD database for AlexNet, ResNet50 and FaceNet. The difference of results between the two databases comes down to the fact that in the SMFD database, mostly, masks are not really well placed on the face, since the masks are synthetic and are not real-world as in the case of the database RMFD database.

Regarding the models performance, FaceNet performed better on both databases. Although, AlexNet and ResNet50 models show significant results in unmasked face recognition, they present a remarkable degradation with masked faces. Even with the adjustment of some parameters for each architecture, as well as with the training of the models by a large number of masked faces.

According the authors in [35], AlexNet model reached higher accuracy of 100% on YTF datasets, 99.55% and 99.17% for the GTAV face and ORL datasets respectively. While ResNet50 has achieved high accuracy of 100% on GTAV face and YTF datasets. Similarly, FaceNet model presents a degradation compared to the results provided with the unmasked faces, let us quote the example of literature [50] where authors mentionned that FaceNet is highly efficient in non-masked faces recognition that it can reach 100% accuracy on YALE, JAFFE, AT&T datasets. Although the masked faces influenced the performance of the FaceNet model, it remains robust relatively to other models.

We have improved the classification process by a combination of voting-based classifiers which greatly improves the performance of the model, where we have reached an accuracy rate equals to 94.46% with RMFRD. This combination has been used in several works for different objects recognition other than masked faces recognition. Let's cite the example of the article [56], where the authors obtained an accuracy rate equals to 97.11% for Arabic-word handwriting recognition. In literature [57], classifiers combination for recognition of Arabic literal provides an accuracy rate equals to 98%.

It is obvious that these values exceed ours, this excess in values comes down to the fact that the models are less complex, the stains are easier and even the databases are much smaller.

To further appraise the proposed method, we have compared our model with several techniques destined specially for recognizing masked faces, such as the model implemented in [58] that just combined FaceNet with SVM and they got an accuracy rate equals to 91.304%. In the study presented in [59], authors used VGG16, the Multilayer Perceptron Classifier (MLP) and Bag-of-Features (BoF) paradigm, the accuracy rate obtained is equals to 91.3% . An efficient face recognition method presented in [35] using Transfer learning (ResNet50 and AlexNet) to fine-tune pre-trained models to the masked

face detection dilemma using an SVM classifier, authors have obtained an accuracy equal to 87%. Table VII summarizes this comparative study.

TABLE VII. PERFORMANCE COMPARISON WITH STATE-OF-THE-ART METHODS

Model	Dataset	Accuracy
FaceNet+Svm	RWMFD	91.304%
VGG16+BOF	RWMFD	91.3%
CNN+Svm	RWMFD	87%
Our model	RWMFD	94.46%

These methods achieve lower values than obtained by the proposed technique, which confirms the effectiveness of the classifiers combination. Then, the complementarity between the two classifiers increases the recognition accuracy and robustness of the model, this comes down to the fact that in image classification, the concept of ambiguity is particularly related to the presence of noisy pixels, mixed pixels and pixels from regions that have undergone changes. If some pixels are between two classes (such as those located on the boundaries of homogeneous regions), those pixels should be classified into a union of two classes rather than a single class. This case can be important when the spatial resolution of the sensor is high. Mixed pixels intervene in the image modeling of a single source whenever that source cannot distinguish between the two classes. In this case, only class related information is available. Pixels in areas with little change are difficult to distinguish from pixels in stable areas and must therefore use two classifiers to get more accurate results.

## VII. CONCLUSION

Today, the task of recognizing a masked face is a challenging process which makes it a focus of interest for scientific committees, given the importance of facial recognition for the security of various organizations and applications around the world. Models that intended for the recognition of unmasked faces have become helpless and unable to provide satisfactory performance to the expectations of security systems and real-time applications. Over the last two years, several techniques have emerged for this purpose but these techniques always remain less effective than the models destined of unmasked faces. This fact comes to several factors. First, the loss of the majority of facial details. Second, all the proposed techniques are based on models previously intended for unmasked faces by adjusting a few parameters to render the model adaptable with the new task. Third factor, the databases designed for the study of masked faces recognition systems, require more improvement. In this regard, we have proposed a model based on FaceNet with combination of classifiers (SVM-KNN) in order to have satisfactory results. This combination gives better results compared to the classification by a single classifier given the complementarity between SVM and K-NN classifiers. Finally, safety is vital at all levels (social, industrial, services, etc.) and security systems must reach a certain level of robustness, that's why our future work aims to develop a new model which does not consider masked faces as a barrier to having excellent results.

## REFERENCES

- [1] Organization WH, et al, "Advice on the use of masks in the context of Covid-19, interim guidance", Tech. rep., World Health Organization, 2022.
- [2] Centers For Disease Control (CDC), "Your health", 2022.
- [3] Ngan, M. L. and Grother, P. J. and Hanaoka, K. K, "Ongoing Face Recognition Vendor Test (FRVT) Part 6A: Face Recognition Accuracy with Masks Using Pre-COVID-19 Algorithms, Internal Report (NIS-TIR), National Institute of Standards and Technology, Gaithersburg, MD, USA, 2020.
- [4] Hariri, W, "Efficient masked face recognition method during the COVID-19 pandemic", Signal Image Video Process, Vol.16(3), pp.605-612, 2022, 10.1007/s11760-021-02050-w.
- [5] Wang, Zhongyuan, Wang, Guangcheng, Huang, Baojin, Xiong, Zhangyang and Hong, Qi and Wu, Hao. et al, "Masked Face Recognition Dataset and Application", arXiv, 2020.
- [6] Ligang Zhang, Brijesh K. Verma, Dian Tjondronegoro and Vinod Chandran, "Facial Expression Analysis under Partial Occlusion", ACM Computing Surveys (CSUR), Vol.51, pp.1-49, 2018.
- [7] Lahasan, B, Lutfi, S.L. and San-Segundo, R. "A survey on techniques to handle face recognition challenges: occlusion, single sample per subject and expression". Artif Intell, Rev 52, pp.949-979, 2019.
- [8] Zeng Dan, Veldhuis Raymond and Spreuwers Luuk, "A survey of face recognition techniques under occlusion", IET Biometrics, Vol. 10(6), pp.581-606, <https://doi.org/10.1049/bme2.12029>, 2021.
- [9] GuiLing, W. "Masked Face Recognition Algorithm for a Contactless Distribution Cabinet", Mathematical Problems in Engineering, 10.1155/2021/5591020, 2021.
- [10] Wang, S., Xia, K., Wang, L., Zhang, J. and Yang, H., "Improved RPCA method via non-convex regularisation for image denoising", IET Signal Process, pp. 269-277, <https://doi.org/10.1049/iet-spr.2019.0365>, 2020.
- [11] Priya. GN. and Banu. RW., "Occlusion invariant face recognition using mean based weight matrix and support vector machine", Sadhana, Vol.39(2), pp.303-315, 10.1007/s12046-013-0216-3, 2014
- [12] Andesh Khomele. V. and Mundada. S. G., "An Approach for Removal of Occlusion using Exemplar based Image-Inpainting Technique: A Review", International Journal of Engineering research & Technology (IJERT), vol.4, 2015.
- [13] Rajeswari.G. and Ithaya Rani. P., "Face Occlusion Removal for Face Recognition Using the Related Face by Structural Similarity Index Measure and Principal Component Analysis", Journal of Intelligent & Fuzzy Systems, Vol.42(6), pp. 5335 - 5350, 2022.
- [14] He, L., Li, H., Zhang, Q. and Sun, Z., "Dynamic feature learning for partial face recognition", The IEEE conference on computer vision and pattern recognition, Vol.42, pp. 7054-7063, 2018.
- [15] Duan, Q and Zhang, L., "Look more into occlusion: Realistic face frontalization and recognition with boostgan", IEEE Transactions on Neural Networks, Vol.42(6), pp. 1-15, 2020.
- [16] Nizam Ud Din, Kamran Javed, Seho Bae and Juneho Yi, "A Novel GAN-Based Network for Unmasking of Masked Face", IEEE Access, Vol.8, pp. 44276-44287, 2020.
- [17] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li and Yu Qiao, "Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks", Institute of Electrical and Electronics Engineers (IEEE), IEEE Signal Processing Letters, Vol.23(10), pp. 1499-1503, 10.1109/lsp.2016.2603342, oct, 2016.
- [18] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Ubowaja, Michael Hays. et al, "MediaPipe: A Framework for Perceiving and Processing Reality", Third Workshop on Computer Vision for AR/VR at IEEE Computer Vision and Pattern Recognition (CVPR), 2019.
- [19] Alzu'bi Ahmad, Albalas Firas, AL-Hadhrami Tawfik, Younis Lojin Bani and Bashayreh Amjad, "Masked Face Recognition Using Deep Learning: A Review", Electronics, Vol.10(21), 2021.
- [20] Yalavarthi Bharat Chandra and G. K. Reddy, "A Comparative Analysis Of Face Recognition Models On Masked Faces", International Journal of Scientific & Technology Research, Vol.9, pp.175-178, 2020.
- [21] Nizam Ud Din, Kamran Javed, Seho Bae and Juneho Yi, "Effective Removal of User-Selected Foreground Object From Facial Images



- Using a Novel GAN-Based Network”, IEEE Access, Vol.8, pp. 109648-109661, 2020.
- [22] Tero Karras, Timo Aila, Samuli Laine and Jaakko Lehtinen, “Progressive Growing of GANs for Improved Quality, Stability, and Variation”, ArXiv, 2018.
- [23] Mengyue Geng, Peixi Peng, Yangru Huang and Yonghong Tian, “Masked Face Recognition with Generative Data Augmentation and Domain Constrained Ranking”, Proceedings of the 28th ACM International Conference on Multimedia, 2020.
- [24] Sengupta Soumyadip, Chen Jun-Cheng, Castillo Carlos, Patel Vishal M, and Chellappa, Rama, Jacobs, David W., “Frontal to profile face verification in the wild”, IEEE Winter Conference on Applications of Computer Vision (WACV), pp.1-9, 10.1109/WACV.2016.7477558, 2016.
- [25] Feifei Ding, Peixi Peng, Yangru Huang, Mengyue Geng, Yonghong Tian, “Masked Face Recognition with Latent Part Detection”, Proceedings of the 28th ACM International Conference on Multimedia, 2020.
- [26] Aqeel Anwar and Arijit Raychowdhury, “Masked Face Recognition for Secure Authentication”, ArXiv, 2020.
- [27] Cao Qiong, Shen Li, Xie Weidi, Parkhi Omkar and Zisserman Andrew, “VGGFace2: A Dataset for Recognising Faces across Pose and Age”, pp. 67-74, 10.1109/FG.2018.00020, 05.2018.
- [28] Deng Hongxia, Feng Zijian, Qian Guanyu, Lv Xindong, Li Haifang and LiGang, “MFCosface: A Masked Face Recognition Algorithm Based on Large Margin Cosine Loss”, Applied Sciences, Vol.11(16), 2021.
- [29] Yuan Xiaowei and Park In Kyu, “Face De-Occlusion Using 3D Morphable Model and Generative Adversarial Network”, IEEE/CVF International Conference on Computer Vision (ICCV), pp. 10061-10070, 10.1109/ICCV.2019.01016, 2019.
- [30] Parkhi O.M., Vedaldi, A. and Zisserman, A., “Deep Face Recognition”, the British Machine Vision Conference (BMVC), Swansea, pp. 41.1-41.12, 2015.
- [31] Yi Dong, Lei Zhen, Liao Shengcai and Li Stan Z., “Learning Face Representation from Scratch”, arXiv, 2014.
- [32] Aaron Nech, Ira Kemelmacher-Shlizerman, “Level Playing Field for Million Scale Face Recognition”, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3406-3415, 2017.
- [33] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton, “ImageNet classification with deep convolutional neural networks”. Communications of the ACM, Vol.60, pp. 84-90, <https://doi.org/10.1145/3065386>, 2017.
- [34] Lu Yang, “Image Classification Algorithm Based on Improved AlexNet in Cloud Computing Environment”, 2020 IEEE International Conference on Industrial Application of Artificial Intelligence (IAAI), pp. 250-253, 10.1109/IAAI51705.2020.9332891, 2020.
- [35] Almabdy Soad and Elrefaei Lamiaa, “Deep Convolutional Neural Network-Based Approaches for Face Recognition”, Applied Sciences, Vol.9(20), 10.3390/app9204397, 2019.
- [36] Simonyan Karen and Zisserman Andrew, “Very Deep Convolutional Networks for Large-Scale Image Recognition”, arXiv, 10.48550/ARXIV.1409.1556, 2014
- [37] Tony Gwyn, Roy Kaushik and Mustafa Atay, “Face Recognition Using Popular Deep Net Architectures: A Brief Comparative Study”, Future Internet, Vol.13, p.164, 2021.
- [38] He, K., Zhang, X., Ren, S. and Sun, J., “Deep Residual Learning for Image Recognition”, The IEEE Conference on Computer Vision and Pattern Recognition”, pp. 770-778, Las Vegas, NV, 27-30 June 2016.
- [39] Tavakolian Niloofar, Nazemi Azadeh, Azimifar Zohreh and Murray Iain, “Face recognition under occlusion for user authentication and invigilation in remotely distributed online assessments”, International Journal of Intelligent Defence Support Systems, Vol.5, p.277, 10.1504/IJIDSS.2018.099889, 12.2018.
- [40] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, et al., “MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications”, ArXiv, Vol. abs/1704.04861, 2017.
- [41] Tony Gwyn, Roy Kaushik and Mustafa Atay, “Face Recognition Using Popular Deep Net Architectures: A Brief Comparative Study”, Future Internet, Vol.13, p.164, 2021.
- [42] Szegedy Christian, Vanhoucke Vincent, Ioffe Sergey, Shlens Jon and Wojna Zbigniew, “Rethinking the Inception Architecture for Computer Vision”, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2818-2826, 10.1109/CVPR.2016.308, 2016.
- [43] Szegedy Christian, Ioffe Sergey, Vanhoucke Vincent and Alemi Alex, “Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning”, arXiv, 10.48550/ARXIV.1602.07261, 2016.
- [44] Chollet François, “Xception: Deep Learning with Depthwise Separable Convolutions”, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1800-1807, 10.1109/CVPR.2017.195, 2017.
- [45] Florian Schroff, Dmitry Kalenichenko and James Philbin, FaceNet: A unified embedding for face recognition and clustering, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, pp. 1-10, 10.1109/cvpr.2015.7298682, 2015.
- [46] Marcelo Beckmann, Nelson F. F. Ebecken and Beatriz S. L. Pires de Lima, “A KNN Undersampling Approach for Data Balancing”, Journal of Intelligent Learning Systems and Applications, Vol.7, November 11, 2015.
- [47] Krizhevsky, A., Sutskever, I. and Hinton, G.E., “ImageNet Classification with Deep Convolutional Neural Networks”, ACM, Vol.60, pp. 84-90, 2017.
- [48] Le Khanh, “A study of face embedding in face recognition”, 2019.
- [49] Gregory R. Koch, “Siamese Neural Networks for One-Shot Image Recognition”, 2015.
- [50] William Ivan, Ignatius Moses Setiadi, De Rosal and Rachmawanto, Eko Hari, Santoso and Heru Agus. et al, “Face Recognition using FaceNet (Survey, Performance Test, and Comparison)”, Fourth International Conference on Informatics and Computing (ICIC), pp. 1-6, 10.1109/ICIC47613.2019.8985786, 2019.
- [51] Jean-Pierre PECUCHET, “Evolutionary Optimisation of Kernel Functions for SVMs”, 2015.
- [52] Wirdiani Ayu, Hridayami Praba, Widiari Ayu, Rismawan Komang, Candradinata Putu and Jayantha, I., “Face Identification Based on K-Nearest Neighbor”. Scientific Journal of Informatics. Vol.6. pp.150-159. 2019.
- [53] M. Mohandes, M. Deriche and S. O. Aliyu, “Classifiers Combination Techniques: A Comprehensive Review”, in IEEE Access, vol. 6, pp. 19626-19639, 2018.
- [54] Ahmad H. A. Eid, “Combined Classifiers for Invariant Face Recognition”, arXiv, 2016
- [55] Li Yixiao, Fang Yuan, Haipeng Peng and Yang Yixian, “Bagged Tree Based Frame-Wise Beforehand Prediction Approach for HEVC Intra-Coding Unit Partitioning”. Electronics. Vol.9. 1523. 10.3390/electronics9091523, 2020.
- [56] Karim Alia and Alawi Mohammed, “Combine SVM and KNN Classifiers for Handwriting Arabic Word Recognition based on Multi feature”, 2019.
- [57] Zaghdoudi Rachid and Seridi Hamid, “Combination of Multiple Classifiers for Off-Line Handwritten Arabic Word Recognition”. International Journal of Information Technology. Vol.14. pp.713-720, 2017.
- [58] Hamdi, Houssamddine and Yurtkan, Kamil, “Masked Face Recognition Based on FaceNet Pre-Trained Model”, 2022.
- [59] walid hari, “Efficient Masked Face Recognition Method”, p.8, 2020.



# Human Position and Object Motion based Spatio-Temporal Analysis for the Recognition of Human Shopping Actions

Nethravathi P. S<sup>1</sup>

Faculty, College of Computer and  
Information Sciences Srinivas  
University  
Mangalore, India

Karuna Pandith<sup>2</sup>

Faculty, Department of Information  
Science & Engineering NMAM  
Institute of Technology  
Nitte, Karnataka, India

Manjula Sanjay Koti<sup>3</sup>

Faculty, Dept. of MCA  
Dayananda Sagar Academy of  
Technology & Management  
Karnataka, India

Rajermani Thinakaran<sup>4\*</sup>

Faculty of Data Science and Information Technology  
INTI International University  
Nilai, Negeri Sembilan, Malaysia

Sumathi Pawar<sup>5</sup>

Faculty, Department of Information Science & Engineering  
NMAM Institute of Technology  
Nitte, Karnataka, India

**Abstract**—Retailers have long sought ways to better understand their consumers' behavior in order to deliver a smooth and enjoyable shopping experience that draws more customers every day and, as a result, optimizes income. By combining various visual clues such as activities, gestures, and facial expressions, humans may fully grasp the behavior of others. However, due to inherent problems as well as extrinsic forced issues such as a shortage of publicly available information and unique environmental variables, empowering computer vision systems to provide it remains an ongoing problem (wild). In this paper, the authors focus on identifying human activity recognition in computer vision, which is the first and by far the most important cue in behavior analysis. To accomplish this, the authors present an approach by integrating human position and object motion in order to detect and classify tasks in both temporal and spatial analysis. On the MERL shopping dataset, the authors get state-of-the-art results and demonstrate the capabilities of the proposed technique.

**Keywords**—Deep convolutional neural networks; computer vision; object detection; object localization; temporal analysis; human shopping actions component

## I. INTRODUCTION

For years, the computer vision industry has been working on recognizing human actions. Many essential applications require the ability to recognize diverse behaviors from video data, such as fight identification from surveillance footage, human-robot interaction, video streaming analysis for online streaming services, and home security surveillance. Action recognition's main purpose is to recognize human actions in a video frame. For video-sharing services like YouTube and Twitch, action recognition is indeed a must-have feature. It can decipher a video's content and determine whether or not it should be made public. This tool can assist in the filtering of potentially harmful videos, such as bomb-making methods, choking activities, and the use of hard narcotics [1-3].

However, in areas like retail and shopping, the impact has been minimal. Using such technology in this context has a number of advantages, including efficient monitoring, consumer behavioral analysis, targeted marketing, and so on. Retailers will benefit from increased efficiency and revenue, as well as a more convenient shopping experience for customers if these strategies are used. Furthermore, the share of the worldwide trade market that these technological solutions occupy might be deduced from the rapidly expanding demand for them. The actual worth of such Artificial Intelligence (AI) based solutions relating to the retail industry is expected to be about US\$10 billion by 2025, according to research conducted by Grand View Research [2].

Applying AI, and particularly machine learning approaches, to the shopping sector is still difficult, due to the insufficiency of data, primarily because of security issues, expensive labeling, as well as the need to stay proprietary where data is gathered. In spite of the datasets being publicly available to the researchers, (e.g., The MERL shopping dataset [4]), applying deep learning techniques to those is difficult as the external challenges posed by distinctive environmental factors like camera view angle, quality of the video, interrelations between the goods and the customers, and high obstruction. However, success of the existing deep learning algorithms can be attributed partially to the utilization of largely available public datasets like ImageNet [5], UCF101 [6], or [4], which allow sophisticated methods with numerous variables to be optimally trained. The completed actions are the major visual clue in understanding human behavior, which when paired with additional indicators like facial expressions tracked over time can also provide detailed behavioral knowledge [1, 7, 8,]. To describe human actions, the initial efforts on action recognition adopted Three-dimensional (3D) models [9, 10]. However, creating a 3D model using videos is time-consuming and costly.

As a consequence, people rather employ global or local representations for action recognition. These methods are known as representation-based methods. Currently, deep networks-based algorithms have indeed been able to attain promising results in action detection, because of the fast evolution of graphics processing units (GPU). The task of activity detection and recognition from shopping surveillance footage inputs is used as the primary topic of this paper. The current study only focuses on categorizing the clipped video into the given activities during recognition, whereas categorization is practiced to a continuous video sequence of multiple activities during detection; i.e., the temporal location at the beginning and length of the actions are also unknown and desired [11].

The main contributions of this paper are

- For a less explored camera view angle, we present an innovative strategy powered by Generative Adversarial Networks (GANs) which uses partial body position in the lack of exact joint locations (top view).
- Using the proposed novel method along with the standard transfer learning, we train and test on the MERL shopping dataset, which has different challenges such as camera angle view, classifications of activity, and limited data for training.
- The authors propose a simple but successful technique for using our action-identifying network as an action detector that identifies and classifies action locations in real-time. This method divides the difficult detection task into identification and detection modules and uses a two-stream network to combine diverse sources in semantic space.
- The authors successfully combine two independent sets of features, one for recognizing or detecting the action, namely human body posture (incomplete) and object-of-interest motion, to direct greater network resources and attention to the most significant signals while ignoring the less important ones. This is performed through the use of self-attention, in which the video's relevant spatial regions are connected to other regions in adjacent frames for enhanced accuracy and/or precision.

The remaining sections of the paper are organized as follows: Section II - consists of the extensive literature survey; Section III - is the detailed explanation of the methodology used in the study; Section IV - is a detailed presentation of the results arrived at along with comparisons. We conclude the paper by providing future directions.

## II. LITERATURE SURVEY

Extraction through feature engineering algorithms that can effectively recognize and depict motion in the input pattern of image frames is the focus of this research. Many approaches of mixing optical flow (OF) and feature matching [12] have been introduced since the early phases of the space-time pyramid [13] until recent years, with the most current ones attempting to replace feature extractors using deep neural networks. However, predicting the flow of optical from succeeding video

frames has proven to be quite efficient. The literature on this subject contains a multitude of ways. Recently, there have been attempts to integrate these techniques with deep neural networks in order to achieve the best-of-both-worlds results, Horn et al. [14]. Many indigenous features have been facing this situation in recent years.

Pose estimation from single red, green, and blue (RGB) images has made substantial progress recently [6, 11, 15, 16, 17,], prompting its use as a high-level feature in movies to effectively represent diverse sorts of activities [18, 19]. Single image estimate is relatively reliable, even the tiniest mistake or disturbance in sequential posture estimation in videos is detrimental to activity interpretation utilizing existing futuristic techniques. With regards to missing joint locations in frames, as we will show factually, this occurs rather frequently irrespective of regular or low demanding settings. As a result, several techniques to utilize this vital information as an additional medium of data in combination with other resources such as optical flow and raw input frame have been presented. There are many alternative ways to take advantage of body posture characteristics while disregarding the faults. Indigenous approaches to cipher consecutive pose data into photos for classifying the actions are defined by some. There are numerous approaches to efficiently incorporate posture estimation into activity understanding in case of good pose prediction (for instance, 3D pose employing motion capture skeletal sensors or depth camera) [20]. However, because of a crowd, low range depth and occlusion, and costs, the use of depth cameras or signal fusion techniques is not viable in the shopping environment.

In computer vision, human-object interaction has long been a concern. However, the majority of the considered interactions revolve around sports [21], cooking [22], or ordinary activities [6, 9], which can sometimes be categorized from single photos [21]. Kim et al. [23] present an effective method for recognizing object-based activities. For action recognition from security cameras, the researchers make use of the graph neural networks for merging the object and human pose data. Their method, however, is largely dependent on the quality of the input posture data.

Multi-stream Convolutional Neural Networks (CNNs) have recently been popular for combining diverse modalities of data before generating decisions [4, 16]. It can be seen in most of the presented systems that one stream is dedicated to temporal understanding (usually utilizing a labeled training data and computed algorithm related to flow of optics [17]), whereas another is exclusively for diving into spatial features in the image.

Many academics have been attributed to the success of Two-dimensional (2D) CNNs in image interpretation to investigate the feasibility of doing so in videos. As a result, numerous ways to widen the convolution in time have been developed [7, 23, 24]. One of the issues with these techniques is that most of these are computationally costly, whilst moderately outperforming 2D CNN counterparts.

Several researchers have rethought their application, inventing sophisticated combinations of 2D and 3D CNNs to achieve the best possible outcomes [25, 26]. Nevertheless, the

work of exploiting pose data is still under-explored to our knowledge. Many other researchers integrate the two works and sort them concurrently [27]. Some techniques solely tackle the temporal detection element of the work, whilst others integrate both activities and sort them at the same time. Some are inspired by object identification techniques and use temporal region proposals [10, 18], while others are known as segmentation.

In the natural language processing (NLP) community, many unique strategies to replicate human attention in CNNs were first proposed on machine translation jobs. Some of them are used in sequential tasks to highlight key input frames [28], while others are used to simply focus on spatial regions of relevance [10, 9, 2]. Wang et al. [29] and combine them into one differentiable peripheral module. Moreover, there are works defining a representative capable of discovering the important areas or frames using sophisticated analytics like Reinforcement Learning (RL) [28]. The 3D modeling method was utilized extensively in early action recognition studies. The walker hierarchical model [9] uses a number of hierarchical levels to depict a person. To recognize pedestrians in a video, [10] employs linked cylinders and their progression. The kernel learned by CNN is visualized and found that the bottom layers learn low-level features while the upper layers learn high-level representations. This demonstrates that convolutional architecture may be utilized to extract features [29, 30].

Videos, unlike photographs, have a dynamic nature. Directly adding temporal features to a convolutional architecture is a typical approach of using deep networks for action recognition. To achieve this, Ji et al. [31] postulated the 3D CNN, which also uses 3D kernels to obtain both spatial and temporal information.

Many researchers have made contributions to the field of temporal information integration in CNNs. In the temporal domain, Ng et al. [30] discover that maximal pooling outperforms average and other pooling approaches. Karpathy et al. [32] present the slow fusion model, a new convolutional architecture that receives video clips and processes them via an identical set of layers (with common parameters) to provide outcomes for fully connected layers. The video description will then be generated from these completely connected layers. Tran et al. [33] combine the concepts of the visual geometry group (VGG) [9], Decaf [14], as well as the 3D CNN [29] to develop a 3-D graphics technique that can build 3D graphics out of 2D images (C3D), a generic video descriptor. They use the Sports-1M [32] dataset to train their network and extract video attributes from such a fully - connected layer. Learning temporal information from uncontrolled input is the purpose of a deep generative network [21, 26]. Xing Yan et al. [34]

present a deep Dyn encoder to capture video dynamics, based on the linear dynamic system modeling approach proposed by Doretto et al. [16]. The Long Short-Term Memory (LSTM) autoencoder model is created using the LSTM [26] cell.

The encoder LSTM and the decoder LSTM make up this model. Goodfellow et al. [35] propose an adversarial network to address the training challenges in deep generative networks. The competition between a generative and a discriminative model is referred to as adversarial. Mathieu et al. [36] use the adversarial principle where a multi-scale convolutional network is trained and highlight the benefits of pooling in a generative model. In this body of work, many databases have been added to aid in the development and testing of algorithms [28, 6, 24]. Kinetics [8], which is called the ImageNet [13] of videos, is among the largest. Only a few of these have untrimmed films that can be detected [22, 27]. More crucially, as previously said, the exclusive characteristic of statistics related to retail is the insufficiency of data on which to test frameworks to address the distinctive difficulties. As per our knowledge, the MERL [4] dataset is the only one available for human shopping actions. All of these aspects are addressed using Human Position and Object Motion based spatiotemporal analysis and are tested on the Recognition of Human Shopping Actions. Further details regarding the same are explained in the subsequent sections.

### III. METHODOLOGY

A GAN consists of two approaches: a generic model that is trained to comprehend the probability distribution of the input data and a discriminative model that seeks to distinguish real input samples from false ones [37]. Backpropagation is used to simultaneously train both models. GANs' success has been seen in their wide range of applications, which range from creative picture generation and super-resolution to semi-supervised classification. Using input pictures and noisy and imprecise joint heat maps, the authors propose a conditional GAN-based technique for regressing the precise location of six body joints. In the retail environment, we commonly encounter top-view (or near-top-view) recorded surveillance cameras, which exacerbate the difficulty of deciphering human activity by imposing extra occlusions of various body parts and components. Because of this distinct and demanding perspective, many deep learning professionals have avoided training on these sorts of input images. The posture estimator system that we implemented in our challenge suffers from the same problem, despite producing state-of-the-art results of different distinct activities that support traditional camera view angles Fig. 1 depicts the challenges encountered by the given strategy in the present working dataset.



Fig. 1. Illustration of the Issues Faced due to the Camera Angles.

To improve posture estimate accuracy, we present the GAN structure. It also ensures that the positions of the joints of interest in the dataset are predicted more accurately. The generator is in charge of learning the conditional probability of the joint locations given to the current noisy heat map that is extracted and the input frame, which is a CNN with an architecture similar to Inception-v4 followed by a multi-layer perceptron (stacked). The same has been summarized in Fig. 2.

In our scenario, only six joints are considered essential in the current shopping environment scenario, which are both the left and right shoulders, elbows, and wrists. To begin, consider human posture as a probabilistic heat map indicating the areas of joints that require better attention. Next, treating it as a high-level feature of the object which is moving in the scene, rather than treating pose data as a separate source of information for defining an action, instead of general features extracted with a massive proportion of unnecessary background data to acquire motion data without considering any other specific motion representation, forcing the deep learning architecture for feature extraction with a greater focus on these spots in each frame of the input images. Then, by implementing the GAN fine-tuning step, we reduce even more noise from the heat maps obtained from the input related to the six key joints and obtain its precise location, which is given to our pose stream network as a replacement for the pose heat map channel. The working of the same has been described in Fig. 3.

Refer to (1)

$$y_i = \frac{1}{c(x)} \sum_{\forall j} f(x_i, x_j) g(x_j) \quad (1)$$

where,  $i$  and  $j$  are indices of the locations in frames over the entire sequence (space-time),  $f$  refers to embedded Gaussian mapping of the input pattern, and  $g$  is a linear mapping function. Here  $x_i$  refers to weight for a single location in space-time input, and  $j$  is the set of all other potential places. The Embedded Gaussian Function is the function we're using here. The number of modules inserted into the network for optimal performance has been empirically determined to be two. The LSTM's hidden state vectors are subjected to the second attention mechanism. It actually assigns a scalar weight to every input frame based on the network's learned relevance. These weights were adopted to the LSTM's hidden feature vector as in practice. At every time the first step (relating to the feature vector of the frame is multiplied by the weight value supplied by the temporal attention), the LSTM has a hidden state. This is done in practice where a single fully-connected layer on the LSTM's hidden states is trained. The first of the two modules, in particular, has played a key role in advising us on how to strengthen our approach. The placement of joints is connected with a higher weight, notably the six joints that make up a part of the posture model, as shown in the representations provided in Fig. 3. By substituting the exact heat map of joint locations with the ones from our generator network, there was even more attention forced which improved our outcomes. In this case, by reducing the noisy heat, we were able to stimulate the concentration of network resources with more assurance maps.

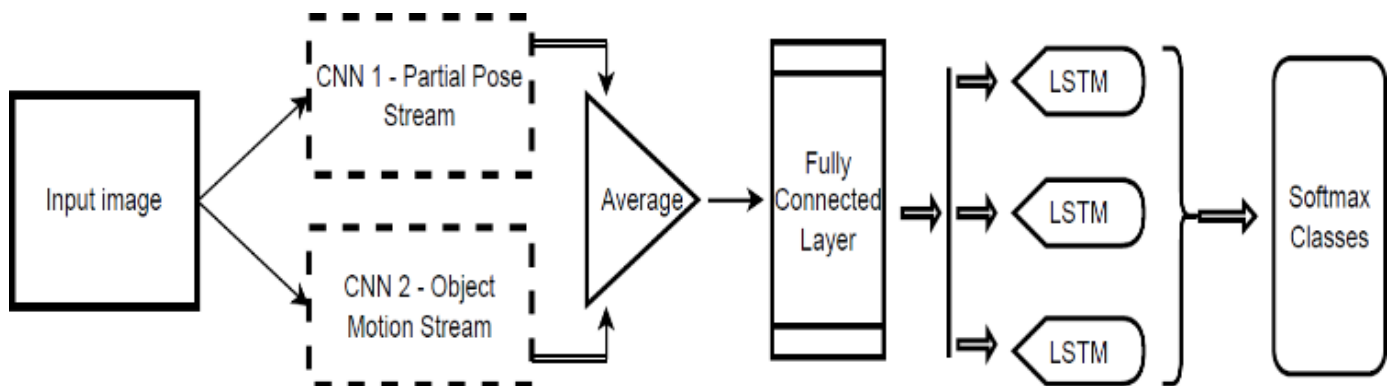


Fig. 2. An Overview of the Proposed GAN Structure.

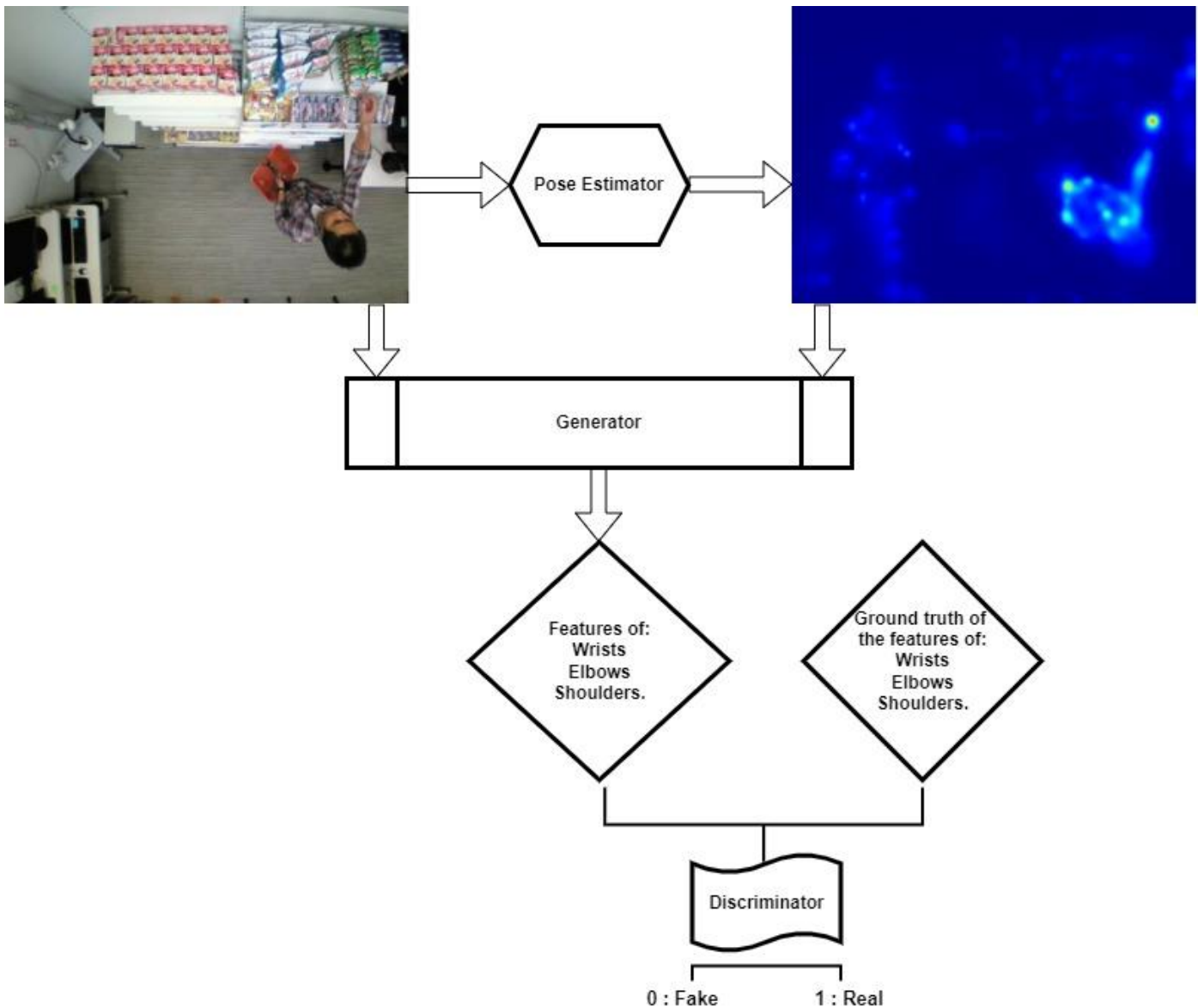


Fig. 3. A Summary of our GAN-Based Pose Fine-Tuning Method.

#### IV. RESULTS AND ANALYSIS

##### A. Experimental Setup

The authors used PyTorch for the implementation as it is open source. The fine-tuned network is then used with the remainder of the network, with no further training, to retrieve earlier data regarding item placements. To construct the weights for the generator in GAN architecture, we manually annotated additional thousand frames which are sampled uniformly with six joint-of-interest (e.g., wrists, elbows, and shoulders) positions. After that, the generator is supervised, and they are pre-trained using the Mean Square Error loss function across the ground-truth joint locations. On the other hand, the discriminator was built up at random using the Xavier approach.

After the GAN training has converged, the generator is used individually along the remainder of the network without

any more fine-tuning, as previously indicated. The pose and object mappings are created using binary maps in the following manner. The map has values of one in a circle of constant radius (here 10 pixels) around each joint center and 0 somewhere else for joint locations, and one in a rectangle area of fixed size (here 40 pixels) around the center of each identified item and zero otherwise for object locations. To reduce the Cross-Entropy loss over the Softmax class probability outputs, the entire recognition network is trained using gradient descent.

Adam is the optimizer that has been proven to perform better in terms of fast convergence in large-scale models like ours. This is achieved by making dynamic (or adaptive, as the authors call it) changes to the learning rate for each weight based on the gradient's higher and lower coefficients so far. The default hyper-parameter settings specified by the authors are = 0.001, 1 = 0.9, 2 = 0.999, and 1008.

It's difficult to train the spatiotemporal attention components in the intermediary layers of the two streams. These modules were given a contribution weight gamma (constant across the network) that was initially set to 0. The authors restart the training for a few epochs (9 to 11) once the main network has converged, starting with 0.1 and gradually increasing it to 1. Finally, because each stream has a similar design to the Inception v4 network, the authors initialize both streams using transfer learning from networks trained on COCO classification tasks. This allows us to drastically minimize training time; the full network training takes only 17-21 epochs to complete.

The detection sliding window and stride parameters must be fine-tuned in the final stage. A Brute Force search in two sets of values is used to accomplish this. As mentioned in the assessment section, the overall performance of each pair of window-stride size values is examined, and the top overall values are picked. Finally, because each stream has a similar architecture to the Inception v4 network, we initialize both streams using transfer learning from networks trained on COCO classification tasks.

This allows us to cut the training time in half; the total network training takes around 17-21 epochs. The detection sliding window and stride parameters must be fine-tuned in the final stage. A Brute Force search in two sets of values is used to accomplish this. As noted in the assessment section, the overall performance of each pair of window-stride size values is evaluated, and the top overall values are picked. Sliding windows range in length from 3 to 45 frames, with a stride of one to the length of the window. The PyTorch deep learning library is used to implement the entire training and inference process in Python.

### B. MERL Dataset Results

The MERL dataset was produced in response to a lack of relevant datasets in retail settings. The six activities are "reach the shelf," "retract from the shelf," "hand in the shelf," "inspect the object," "inspect the shelf," and the backdrop (or no action) class. It was photographed using a roof camera to look like a real retail mall, but not in any detail. There are 42 people in this dataset who work as shoppers. One of the dataset's challenges is that participants must complete a sequence of tasks, which can be easily exploited as a well-behaved transition probability matrix to produce good results on this dataset at the cost of simplifying it as a well-behaved transition probability matrix. Table I is an example of this challenge. This supplies the network with useful prior knowledge that may be used during inference to improve recognition accuracy.

TABLE I. MERL DATASET, ACTIONS, AND ITS CORRESPONDING DISTRIBUTIONS

Actions	Reach	Retract	Hand in	Inp. Product	Insp. Shelf
Reach	0.0	63.8	34.1	0.7	1.4
Retract	21.2	0.0	0.8	49.53	28.47
Hand in	0.12	85.7	0.0	6.32	7.86
Inp. Product	61.08	3.1	0.84	0.0	34.98
Insp. Shelf	98.9	0.094	0.67	0.34	0.0

Nonetheless, one of the key benefits of our technique is that we may get cutting-edge findings without relying on this extensive historical knowledge. This makes sense since, in real-world solutions, powerful priors are difficult to come by and impractical, therefore they can't be used. As a result, the authors eliminated these biases by ensuring that the participants do not follow a straightforward sequence of activities during the testing.

Unlike prior approaches that reported on the MERL dataset, we present results for both recognition and detection. The former believes that the input films only include single action, but the latter receives an untrimmed video with numerous actions as input. Table II displays our MERL recognition results, and Table III compares our detection results to those previously published because the recognition results for this dataset haven't been disclosed before by any other method, this presents a new challenge for future research, allowing other algorithms to compare their detection accuracy. As we've seen and as our results show, the accuracy of detection should be near to the precision of recognition for a decently good detection approach. As compared to other approaches, we significantly have higher Intersection over Union (IoU) (average 0.77 compared to 0.5 as the maximum reported) over previous methods while attaining better detection results, too.

TABLE II. RESULTS OF PROPOSED METHOD ON MERL DATASET

Details	In Percentage
Overall recognition Accuracy	71.14
F1 @ 50	67.11
Frame wise accuracy	71.41

TABLE III. COMPARISON OF RESULTS WITH PROPOSED METHOD (MERL DATASET)

Methodology	F1 {IoU = 0.5}	Accuracy in Percentage (Frame-Wise)
Multi-stream bi-directional RNN [4]	65.4	76.3
Temporal Convolutional Networks	72.9	79.0
Two stream CNN	74.8	77.1
Proposed methodology	77.46	75.13

### V. CONCLUSIONS

The authors have developed a framework for fine-grained activity recognition and detection in retail environments. Due to the short time between each action, considerable intra-class variance, and low inter-class variation, fine-grained detection is difficult; these difficulties contribute to the task's intrinsic complexity. Furthermore, the work is made more difficult by the extrinsic difficulty of a rare and unusual camera viewing angle. We developed a semi-supervised technique employing GAN to fine-tune posture estimate results when there is a discrepancy in the images present at different angles during training and prediction. We gave detailed experimental data to demonstrate the method's applicability in real-world scenarios, particularly in shopping contexts, which have their own set of characteristics and obstacles.



When direct estimation fails, we speculate that combining pose estimation with image sequences is one way to use the associated range of motion as background information for forecasting joint location in the next time step (e.g., occlusion). Furthermore, we believe that describing the attention process as a salient region localization job for an expert system might lower the computation cost of its training and prediction, given the recent success of deep reinforcement learning methodologies. Further attempts at face expression recognition and/or eye gaze prediction provide more data for deep customer behavior analysis.

#### REFERENCES

- [1] K. Soomro, H. Idrees, and M. Shah. Online Localization and Prediction of Actions and Interactions. volume 41, Feb.2019.
- [2] H. B. Zhang, Y. X. Zhang, B. Zhong, Q. Lei, L. Yang, J. X. Du, and D. S. Chen, "A comprehensive survey of vision-based human action recognition methods," *Sensors*, 19(5), pp. 1005, 2019.
- [3] G. Cheng, Y. Wan, A. N. Saudagar, K. Namuduri, and B. P. Buckles, "Advances in human action recognition: A survey," 2015, arXiv preprint arXiv:1501.05964.
- [4] B. Singh, T. K. Marks, M. Jones, O. Tuzel, and M. Shao, "A multi-stream bi-directional recurrent neural network for fine-grained action detection," In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.1961–1970, June 2016.
- [5] J. Deng, W. Dong, R. Socher, L. Li, and and. Imagenet: A large-scale hierarchical image database. In 2009 IEEE Conference on Computer Vision and Pattern Recognition, pages 248–255, June 2009.
- [6] K. Soomro, A. R. Zamir, and M. Shah. UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild.Dec. 2012. arXiv: 1212.0402.
- [7] G. Tripathi, K. Singh, and D. K. Vishwakarma. Convolutional neural networks for crowd behaviour analysis: a survey. pages 1–24. Springer, 2018.
- [8] S. Nigam, R. Singh, and A. K. Misra. A Review of Computational Approaches for Human Behavior Detection. May 2018.
- [9] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu. Spatial Transformer Networks. June. 2015. arXiv: 1506.02025.
- [10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention Is All You Need. June 2017. arXiv: 1706.03762.
- [11] R. Bai, Q. Zhaoy, S. Zhou, Y. Liz, X. Zhaox, and J. Wang. Continuous action recognition and segmentation in untrimmed videos. August 2018.
- [12] H. Wang and C. Schmid. Action recognition with improved trajectories. In 2013 IEEE International Conference on Computer Vision, pages 3551–3558, Dec 2013.
- [13] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In 2008 IEEE Conference on Computer Vision and Pattern Recognition, pages 1–8, June 2008.
- [14] B. K. Horn and B. G. Schunck. Determining optical flow. volume 17, pages 185–203, Aug. 1981.
- [15] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. Convolutional Pose Machines. Jan. 2016. arXiv: 1602.00134.
- [16] A. Newell, K. Yang, and J. Deng. Stacked Hourglass Networks for Human Pose Estimation. In B. Leibe, J. Matas, N. Sebe, and M. Welling, editors, *Computer Vision ECCV 2016*, pages 483–499, Cham, 2016. Springer International Publishing.
- [17] R. A. Gler, N. Neverova, and I. Kokkinos. DensePose: Dense Human Pose Estimation In The Wild. Feb. 2018. arXiv:1802.00434.
- [18] D. C. Luvizon, D. Picard, and H. Tabia. 2d/3d Pose Estimation and Action Recognition using Multitask Deep Learning. Feb. 2018. arXiv: 1802.09232.
- [19] R. Poppe, "A survey on vision-based human action recognition," *Image and vision computing*, 28(6), pp. 976-990, 2010
- [20] R. Vemulapalli, F. Arrate, and R. Chellappa. Human Action Recognition by Representing 3d Skeletons as Points in a Lie Group. pages 588–595, 2014.
- [21] B. Yao and L. Fei-Fei. Modeling mutual context of object and human pose in human-object interaction activities. In 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pages 17–24, June 2010.
- [22] M. Rohrbach, A. Rohrbach, M. Regneri, S. Amin, M. Andriluka, M. Pinkal, and B. Schiele. Recognizing Fine-Grained and Composite Activities using Hand-Centric Features and Script Data. volume 119, pages 346–373, Sept.2016. arXiv: 1502.06648.
- [23] S. Kim, K. Yun, J. Park, and J. Y. Choi. Skeleton-based action recognition of people handling objects. 2019.
- [24] R. Hou, C. Chen, and M. Shah. An end-to-end 3d convolutional neural network for action detection and segmentation in videos. 2017.
- [25] S. Xie, C. Sun, J. Huang, Z. Tu, and K. Murphy. Rethinking Spatiotemporal Feature Learning: Speed-Accuracy Tradeoffs in Video Classification. Dec. 2017. arXiv: 1712.04851.
- [26] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox. FlowNet 2.0: Evolution of Optical Flow Estimation with Deep Networks. Dec. 2016. arXiv: 1612.01925.
- [27] Y. Zhao, Y. Xiong, L. Wang, Z. Wu, X. Tang, and D. Lin. Temporal Action Detection With Structured Segment Networks. pages 2914–2923, 2017.
- [28] Y. Rao, J. Lu, and J. Zhou. Attention-Aware Deep Reinforcement Learning for Video Face Recognition. In 2017 IEEE International Conference on Computer Vision (ICCV), pages 3951–3960, Venice, Oct. 2017. IEEE.
- [29] X. Wang, R. Girshick, A. Gupta, and K. He. Non-local neural networks. IEEE, Jun 2018.
- [30] Joe Yue-Hei Ng, Matthew Hausknecht, Sudheendra Vijayanarasimhan, Oriol Vinyals, Rajat Monga, and George Toderici. "Beyond short snippets: Deep networks for video classification". In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2015, pp. 4694–4702.
- [31] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. "3D convolutional neural networks for human action recognition". In: IEEE transactions on pattern analysis and machine intelligence 35.1 (2013), pp. 221–231.
- [32] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Suk-thankar, and Li Fei-Fei. "Large-scale video classification with convolutional neural networks". In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 2014, pp. 1725–1732.
- [33] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. "Learning spatiotemporal features with 3d convolutional networks". In: Proceedings of the IEEE International Conference on Computer Vision. 2015, pp. 4489–4497.
- [34] Xing Yan, Hong Chang, Shiguang Shan, and Xilin Chen. "Modeling video dynamics with deep dynencoder". In: European Conference on Computer Vision. Springer. 2014, pp. 215–230.
- [35] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. "Generative adversarial nets". In: Advances in neural information processing systems. 2014, pp. 2672–2680.
- [36] Michael Mathieu, Camille Couprie, and Yann LeCun. "Deep multi-scale video prediction beyond mean square error". In: arXiv preprint arXiv:1511.05440 (2015).
- [37] Mogren, O. (2016). C-RNN-GAN: Continuous recurrent neural networks with adversarial training. arXiv preprint arXiv:1611.09904.